The Coalescence of Intrahost HIV Lineages Under Symmetric CTL Attack

Sivan Leviyang

Received: 2 August 2011 / Accepted: 4 May 2012 / Published online: 30 May 2012 © Society for Mathematical Biology 2012

Abstract Cytotoxic T lymphocytes (CTLs) are immune system cells that are thought to play an important role in controlling HIV infection. We develop a stochastic ODE model of HIV–CTL interaction that extends current deterministic ODE models. Based on this stochastic model, we consider the effect of CTL attack on intrahost HIV lineages assuming that CTLs attack several epitopes with equal strength. In this setting, we introduce a limiting version of our stochastic ODE under which we show that the coalescence of HIV lineages can be described through Poisson–Dirichlet distributions. Through numerical experiments, we show that our results under the limiting stochastic ODE accurately reflect HIV lineages under CTL attack when the HIV population size is on the low end of its hypothesized range. Current techniques of HIV lineage construction depend on the Kingman coalescent. Our results give an explicit connection between CTL attack and HIV lineages.

1 Introduction

Cytotoxic T lymphocytes (CTLs) are immune system cells that kill pathogen infected host cells. In the context of HIV infection, considerable experimental evidence suggests that CTLs play a central role in controlling infection and shaping HIV diversity, e.g., Borrow et al. (1994), Carrington and O’Brien (2003), Goonetilleke et al. (2009), Koup et al. (1994), Schmitz et al. (1999).

Roughly, when HIV enters a host cell, typically a CD4+ cell, certain mechanisms within the cell cut up HIV proteins into small pieces (usually 8–11 amino acids long) and present these peptides on the surface of the cell in the form a peptide-MHC complex (pMHC) (DeFranco et al. 2007). CTLs can bind to pMHC complexes and then destroy the presenting cell, but critically each CTL possesses receptors that can
bind to a limited pattern of peptides. An HIV peptide that is attacked by CTLs is referred to as an epitope.

When CTLs attack a given epitope, HIV infected cells possessing that epitope are killed off. However, due to mutation, variants of HIV that do not possess the epitope may exist. Such variants, which are at a selective advantage due to the CTL attack, will proliferate and come to dominate the HIV population. This hypothetical picture has been confirmed in many experimental HIV studies, e.g., Kelleher et al. (2001). Yet despite the putative role of CTLs in controlling HIV infection and the corresponding importance of HIV genetic diversity in evading CTL attack, the impact of CTL attack on intrahost HIV genetic diversity is not well understood.

Most current theoretical tools used in HIV research do not link CTL models to HIV genetic diversity. On one hand, HIV–CTL interaction has been modeled since the beginning of the HIV epidemic (see Nowak and May 2000; Perelson 2002 for a review). Various models are possible, but the standard model consists of a deterministic ODE composed of variables for the population size of HIV virions, infected and uninfected CD4\(^+\) cells, and CTLs targeting infected CD4\(^+\) cells. While the standard model and its many variations give a dynamic picture of HIV and CTL population sizes, they do not connect CTL attack to HIV population genetics.

On the other hand, tools from population genetics that do not explicitly model CTL attack have been applied to HIV. Rodrigo and coworkers used variants of the Kingman coalescent to explore the HIV life cycle and construct inference algorithms based on HIV genetic samples (Rodrigo and Felsenstein 1999; Rodrigo et al. 1999; Drummond and Rodrigo 2000). The popular programs BEAST and LAMARC, which are used to make statistical inferences based on HIV genetic data, assume a Kingman coalescent (Drummond and Rambaut 2007; Kuhner 2006).

In this work, we present results that connect an ODE model of HIV population dynamics under CTL attack to HIV population genetics. More specifically, we consider a stochastic ODE that models HIV escape from CTL attack at multiple epitopes sometime during the chronic phase of infection. Our stochastic ODE describes the dynamics of the HIV population in terms of discrete birth, death, and mutation events, allowing us to specify lineages once the dynamics are given. We show that under a certain small population limit, our stochastic ODE connects to the deterministic ODE models described above.

To connect to HIV population genetics, we consider a collection of HIV-infected cells sampled after HIV has escaped CTL attack. Given a realization of the stochastic ODE dynamics, the lineages of these infected cells can be traced back to the time at which CTL attack initiates, thereby forming a genealogy. For simplicity, we assume CTL attack of equal strength at each considered epitope, a situation we refer to as symmetric attack. In this setting, our main result characterizes the state of the genealogy at the time when CTL attack initiates. Further, we show that HIV escape mutations produce significant stochasticity in the HIV population dynamics.

We analyze our stochastic ODE using methods similar to those used by Iwasa et al. (2005) and Durrett et al. (2009) in their study of cancer pathways. Hermisson and Pennings (2005), Pennings and Hermisson (2006) also used similar techniques in an abstract setting applicable to HIV. Desai and Fisher (2007) analyzed the dynamics of sequential selective sweeps, a model similar in spirit to our own, but they did not allow for varying population size, nor did they construct genealogies. In all these works
and our own, the dynamics of mutations present at low levels in the overall population are well approximated by branching processes. Rouzine and Coffin considered an HIV model that bears some similarity to our HIV–CTL model (Rouzine and Coffin 2010), but their analysis and goals differ from ours.

Our lineage construction is similar in spirit to that of several authors, but there are significant differences between our underlying model and that of previous authors. Kaplan et al. (1988), Durrett and Schweinsberg (2004) considered lineages from a population that has undergone a strong selective sweep, while Barton et al. (2004) considered lineages from a population under selection–mutation equilibrium. Both these works considered a fixed size, Moran model with weak mutation rates. In our case, the stochastic ODE considered does not assume a fixed population size, and we consider a strong mutation rate reflective of HIV biology.

In Sect. 2 we describe our model. In Sect. 3 we describe our theoretical results along with associated numerical results. In Sect. 4 we discuss some implications of our results. Sections 5 and 6 provide proofs of the results presented in Sect. 3. In these two sections, we have endeavored to focus on the intuition behind the proofs. Our hope is that the mathematics presented in these sections contributes to intuition and biological motivation. We place arguments that are mathematically technical, and unnecessary for intuition, in the Appendix.

2 A Model of HIV Dynamics Under CTL Attack

To specify our model, in Sect. 2.1, we introduce terminology that will help characterize the CTL attack. In Sect. 2.2, we introduce our stochastic ODE model and connect it to a deterministic ODE similar to those mentioned in the introduction. In Sect. 2.3, we specify a specific parameter choice for our stochastic ODE that models symmetric CTL attack. Finally, in Sect. 2.4, we discuss genealogies within the context of our HIV–CTL model.

2.1 Escape Graph

We model an HIV population exposed to attack at \( e \) epitopes. To do this, we categorize the HIV infected cells by the presence, represented by a 0, or absence, represented by a 1, of a given epitope. Since there are \( e \) epitopes, the different HIV infected cell variants can be associated with a binary number of length \( e \). For example, if \( e = 2 \), then the HIV infected cell variant, hereafter we simply say variant, \( 10 \) represents an infected cell containing only the second epitope. Intuitively, we think of 1s as representing mutations that alter a locus on the HIV genome that is responsible for producing the attacked epitope.

We let \( E \) be the set of possible variants. In this work, we focus on two possible choices for \( E \). In the first, which we label as \( E_{\text{full}} \), we consider every possible combination of epitopes. For example, if \( e = 3 \), we define

\[
E_{\text{full}} = \{000, 001, 010, 011, 100, 101, 110, 111\}.
\]
In the second choice, which we label as $\mathcal{E}_{\text{linear}}$, we consider only variants that, from left to right, contain a sequence of all $1$s followed by a sequence of all $0$s. In the case $e = 3$, we define

$$\mathcal{E}_{\text{linear}} = \{000, 100, 110, 111\}. \quad (2)$$

Given $\mathcal{E}$, we define a graph $G$ which we call the escape graph of $\mathcal{E}$. $G$ is formed from vertices labeled by elements of $\mathcal{E}$ and arrows that connect a vertex with label $v'$ to one with label $v$ if a single epitope mutation can change $v'$ to $v$. When $\mathcal{E} = \mathcal{E}_{\text{full}}$ and $\mathcal{E} = \mathcal{E}_{\text{linear}}$, we refer to the associated $G$ as the full and linear escape graph, respectively. Figures 1 and 2 show the full and linear escape graph, respectively, in the case $e = 3$.

For any $v \in \mathcal{E}$, $\mathcal{P}(v)$ is the set of elements in $\mathcal{E}$ that can be changed into $v$ by transforming a single $0$ into a $1$. Intuitively, we think of $\mathcal{P}(v)$ as the variants that can be transformed into $v$ by a single mutation, and $\mathcal{P}$ stands for parents. To be clear, if $v = 110$, then we have $\mathcal{P}(v) = \{010, 100\}$ and $\mathcal{P}(v) = \{100\}$ for the full and linear escape graphs, respectively.

We say that the HIV population has escaped CTL attack when all infected cells are of type $111 \ldots 1$. In other words, mutations that remove each of the attacked epitopes have fixed in the HIV population.

### 2.2 ODE

Let $h$ represent the number of uninfected CD4+ cells that are targets for HIV infection. For each $v \in \mathcal{E}$, let $e_v$ be the number of CD4+ cells infected by $v$ variants. We assume birth and death rates for $h$ and the $e_v$ as specified in Table 1. $\lambda$ represents the total birth rate, and $g$ represents the death rate per cell, respectively, of a CD4+ cell in the absence of HIV infection. $b_v h$ and $k_v$ are the birth and death rates per cell of a $v$ variant infected CD4+ cell. An infected CD4+ birth event corresponds to an uninfected CD4+ death event, so uninfected CD4+ cells have an additional death term beyond $g$. 

---

Fig. 1  Full escape graph for $e = 3$

Fig. 2  Linear escape graph for $e = 3$
Table 1  Birth, death, and mutation rates

| Cell type (# of cells) | Birth rate | Death rate | Mutation event |
|------------------------|------------|------------|----------------|
| Uninfected \((h)\)     | \(\lambda\) | \((g + \sum_{v \in \mathcal{E}} b_v e_v)h\) | \(-\)          |
| \(v\) infected \((e_v)\) | \(b_v h e_v\) | \(k_v e_v\) | \(\mu \sum_{v' \in \mathcal{P}(v)} b_{v'} e_{v'} h\) |

We let \(\mu\) be the rate per infection event at which mutations occur that remove any one of the \(e\) epitopes. Correspondingly, new \(v\) variants arise from mutations in \(v' \in \mathcal{P}(v)\) with a rate given in the “mutation event” rate column in Table 1. Notice that a “mutation event” in Table 1 refers to the creation of a \(v\) variant from a mutation in some variant contained in \(\mathcal{P}(v)\), not the mutation of a \(v\) variant itself.

Define \(P(f(t))\) to be a Poisson process with jump rate \(f(t)\) at time \(t\). Then, given the rates in Table 1, we have the following stochastic ODE:

\[
\begin{align*}
\frac{dh}{dt} &= dP(\lambda) - dP(gh) - \sum_{v' \in \mathcal{E}} dP(b_{v'} e_{v'} h), \\
\frac{de_v}{dt} &= dP(b_v e_v h) - dP(k_v e_v) + \sum_{v' \in \mathcal{P}(v)} dP(\mu b_{v'} e_{v'} h),
\end{align*}
\]

where the second equation directly above applies for all \(v \in \mathcal{E}\). Each \(P\) in (3) represents an independent Poisson process run at the specified rate; to avoid cumbersome notation, we do not use a distinct notation for each of these processes. In (3) and throughout this work, we ignore back mutations, a mutation from a variant to a less fit variant. Ignoring such mutations does not affect our results.

To make our system variables \((h\) and the \(e_v)\) \(\mathcal{O}(1)\), we rescale \(h\) and each \(e_v\) by \(\mathbb{H}\) and \(\mathbb{E}\), respectively. Intuitively, \(\mathbb{H}\) and \(\mathbb{E}\) correspond to the order at which uninfected but infectable CD4\(^+\) cells and infected, activated CD4\(^+\) cells capable of producing virions exist during HIV infection, respectively. We set \(\mathbb{H} = \lambda / g\), the steady state of uninfected cells in the absence of HIV infection. This scaling is supported by empirical results suggesting that, at least prior to AIDS onset, the number of uninfected CD4\(^+\) cells during and prior to HIV infection are on the same order (Crandall 1999).

Without justification for a moment, we choose \(\mathbb{E} = g / \bar{b}\) where \(\bar{b}\) is on the order of the \(b_v\). If we rewrite \(h\) and \(e_v\) as \(h / \mathbb{H}\) and \(e_v / \mathbb{E}\), respectively, in (3), we arrive at the rescaled system

\[
\begin{align*}
\frac{dh}{dt} &= \frac{g}{\lambda} \left( dP(\lambda) - dP(\lambda h) - \sum_{v' \in \mathcal{E}} dP\left( \frac{b_{v'}}{b} e_{v'} h \right) \right), \\
\frac{de_v}{dt} &= \frac{\bar{b}}{g} \left( dP\left( \frac{b_v}{b} e_v h \right) - dP\left( \frac{k_v g}{b} e_v \right) + \sum_{v' \in \mathcal{P}(v)} dP\left( \mu b_{v'} e_{v'} h \right) \right).
\end{align*}
\]

We would like to recover a deterministic ODE from (4), in this way showing that our present model is an extension of current deterministic models. Kurtz (1981) showed that one can recover deterministic population ODEs by taking large population limits of stochastic population ODEs. In our context, we can consider \(\mathbb{H} \to \infty\).
and $E \to \infty$. Such limits are reasonable for HIV due to its enormous population size, but it is not immediately clear what the relationship should be between $H$ and $E$ as both go to infinity.

To address this issue in a simple context, consider (4) without CTL attack. In this setting, we need not distinguish between different variants, reducing our system to the variables $h$ and $e$, and we may also ignore mutation. Taking $H$ and $E$ large, we can largely ignore the stochasticity of (3) and arrive at the following deterministic ODE:

$$\frac{dh}{dt} = g(1 - h - eh),$$

$$\frac{de}{dt} = \frac{b}{g} e \left( \lambda h - \frac{kg}{b} \right),$$

which has the equilibrium

$$h = \frac{(kg)}{(b\lambda)}. \quad (6)$$

Consider the variables in the expression for $h$ directly above. The death rate of infected CD4$^+$ cells, $k$, has been measured at approximately two days (Perelson et al. 1996). If we take two days as our time scale, we then expect $k \approx 1$. Uninfected CD4$^+$ cells last on the order two weeks, giving $g = 0.1$ as a reasonable choice. Estimates for $b$ and $\lambda$ have significant variation in the literature. However, we can understand the role of $\lambda$ and $b$ in (4) by noting the following relation:

$$\lambda b = g^2 \left( \frac{H}{E} \right), \quad (7)$$

which follows from our formulas for $H$ and $E$. The above relation and (6) demonstrate that (4) only converges to a deterministic system in the large population limit of $H, E \to \infty$ if the ratio $H/E$ converges to a fixed constant.

To force (4) to have a deterministic limit, we introduce a parameter $\gamma = \lambda \tilde{b}/g$ or, in terms of $H$ and $E$, $\gamma = g(H/E)$ (the factor $g$ is not essential, but gives the system directly below a simpler form). From a biological point of view, $\gamma$ is an inverse measure of the fraction of infectable cells that are actually infected. Empirical results for the ratio of infected to infectable cells are difficult as most infected CD4$^+$ are in the lymph nodes and many such CD4$^+$ are infected but inactive (Levy 1998). However, estimates in the range of 0.01 to 0.1 have been given by several authors and seem reasonable (Levy 1998). With $g = 0.1$, the corresponding range for $\gamma$ is 1 to 10. Using this scaling of $\gamma$, our definition of $E$ is justified biologically.

Rewriting (4) using $\gamma$ and setting $\tilde{b}_v = b_v/\tilde{b}$ gives

$$dh = \frac{1}{E} \left( \frac{g}{\gamma} \right) \left( dP(\gamma E) - dP(\gamma EH) - \sum_{v' \in E} dP(\tilde{b}_{v'} \gamma E e_{v'} h) \right),$$

$$de_v = \frac{1}{E} \left[ dP(\gamma \tilde{b}_v E e_v h) - dP(k_v E e_v) + \sum_{v' \in P(v)} dP(\mu \tilde{b}_{v'} \gamma E e_{v'} h) \right],$$

and we consider the limit of this system as $E \to \infty$. 

Springer
While (8) approaches a deterministic limit as $E \to \infty$ when mutation is ignored, the system will continue to be stochastic if $\mu$ is sufficiently large with respect to $E$. Indeed, as we show in Sect. 3, the scaling of $\mu$ that produces stochasticity is precisely a scaling in which HIV lives. Roughly, stochasticity of (8) exists even as $E \to \infty$ because the dynamics of variants that are of scaled population size $O\left(\frac{1}{E}\right)$, i.e., $e_v = O\left(\frac{1}{E}\right)$, will be stochastic even as $E \to \infty$.

However, if we ignore mutation, then as $E \to \infty$, (8) becomes

$$
\frac{dh}{dt} = g\left(1 - h - \sum_{v' \in E} \tilde{b}_{v'} e_{v'} h\right),
$$

$$
\frac{de_v}{dt} = e_v (\tilde{b}_v \gamma h - k_v),
$$

which has the form of a predator–prey system. Equation (9) is a simplified version of the standard deterministic ODE used today for HIV modeling, the full version includes the virion population. But generally, the reduction to (9) demonstrates how (8) is based on current HIV models.

2.3 Symmetric CTL Attack and Initial Conditions

We consider (8) restricted to the case of symmetric attack. To make the notion of symmetric attack precise, we partition the collection of variants, $E$, into subsets $E_i$ such that

$$
E_i = \{v \in E : v \text{ has } i \text{ 1s in its binary expression}\}
$$

For example, if $e = 4$, then $E_1 = \{1000, 0100, 0010, 0001\}$ and $E_1 = \{1000\}$ for the full and linear escape graph, respectively. $E_i$ is the collection of variants that are mutated at $i$ epitopes. We refer to the $E_i$ generally as variant classes and $E_i$ specifically as the $i$th variant class.

To model symmetric attack, we assume that a variant $v \in E_i$ will be exposed to CTL attack at $e - i$ epitopes, and we assume that the death rate due to CTLs at each single epitope has rate $\Delta k$. We scale time so that infected cells die, in the absence of CTL attack, at rate $1$. As mentioned, the lifetime of an infected cell has been shown to be approximately two days, which is in turn our unit of time. All this is made precise by taking the death rate $k_v$ of $v \in E_i$ to be given by

$$
k_v = 1 + (e - i) \Delta k.
$$

Finally, as an added simplification, we take $b_v$ to be constant. In (8), this amounts to taking $\tilde{b}_v = 1$ by setting $\bar{b} = b_v$.

Before presenting our final system, we note that mutations do not play a role in the equation for $h$ and so stochastic effects will have little impact on $h$ dynamics. For simplicity, and with error that goes to 0 as $E \to \infty$, we replace the $h$ equation by its deterministic counterpart. Putting all these remarks together, we arrive at the following system:
\[
\frac{dh}{dt} = g \left( 1 - h - \sum_{v' \in E} e_{v'h} \right),
\]
\[
d_e(v) = \frac{1}{e} \left( dP(\gamma e_v h) - dP(k_{v'} e_{v'} h) + \sum_{v' \in \mathcal{P}(v)} dP(\mu \gamma e_{v'h}) \right).
\]  

From this point on, we take (12) as describing the dynamics of the HIV population. To set initial conditions, we assume that variant \( v_0 = 0\ldots0 \) is the dominant variant prior to CTL attack. Indeed, CTLs proliferate in response to epitopes existing in the population, so taking \( v_0 \) to be the dominant variant is biologically reasonable. Ignoring other variants for a moment, we set \( h, e_{v_0} \) at time \( t = 0 \) according to the equilibrium of (5):

\[
h(0) = \frac{1}{\gamma},
\]
\[
e_{v_0}(0) = \frac{1 - h(0)}{h(0)}.
\]

Prior to CTL attack, we assume that other variants are at a slight fitness disadvantage to \( v_0 \) and arise due to mutations on \( v_0 \) variants. Assuming, as we have just done, that there are \( O(\mathbb{E}) \) \( v_0 \) variants, there will be \( O(\mu \mathbb{E}) \) variants from each of the \( \mathcal{E}_1 \) classes. From this we can conclude that the number of variants in the \( \mathcal{E}_2 \) class will be of order \( O(\mu^2 \mathbb{E}) \). As we mention below, \( \mu^2 \mathbb{E} \approx 0 \), and so we assume that no \( \mathcal{E}_i \) variants exist at \( t = 0 \) for \( i > 1 \). For simplicity, we assume that \( e_{v}(0) = \mu \mathbb{E} \) for all \( v \in \mathcal{E}_1 \).

These initial conditions are not essential to our results, other choices are possible. Which initial conditions are appropriate will depend on the period of HIV infection one has in mind. We have made a specific choice for the sake of clarity.

2.4 Genealogies

When all variants are of type \( 1\ldots1 \), the HIV population has escaped CTL attack. We let \( T_{\text{sample}} \) be a time after such an escape has been completed and consider \( n \) infected cells sampled at \( T_{\text{sample}} \). Since (12) defines discrete birth and death events, we can construct lineages corresponding to the ancestral lines of these \( n \) sampled cells.

We label the lineages \( \ell_1, \ell_2, \ldots, \ell_n \), and we let \( \Pi(t) \) represent the partition structure of the lineages at time \( t \). To explain this, consider Fig. 3 which represents a possible lineage structure for the case \( n = 8 \). The values of \( \Pi(t) \) at \( t = T_{\text{sample}}, T_B, T_C \) are given by

\[
\Pi(T_{\text{sample}}) = \{ \{\ell_1\}, \{\ell_2\}, \ldots, \{\ell_8\} \},
\]
\[
\Pi(T_B) = \{ \{\ell_1, \ell_2\}, \{\ell_3, \ell_4\}, \{\ell_5, \ell_6\}, \{\ell_7, \ell_8\} \},
\]
\[
\Pi(T_C) = \{ \{\ell_1, \ell_2, \ell_3, \ell_4, \ell_5, \ell_6\}, \{\ell_7, \ell_8\} \}.
\]

At time \( T_{\text{sample}} \), all lineages are separate, and \( \Pi(T_{\text{sample}}) \) consequently partitions each lineage into its own set. By time \( T_B \), the pairs of lineages 1 and 2, 3 and 4, 5 and 6,
and 7 and 8 have coalesced. $\Pi(T_B)$ partitions these pairs to reflect this structure. Finally, by time $T_C$, lineages 1 through 6 have coalesced as has the pair 7 and 8. $\Pi(T_C)$ partitions the lineages accordingly.

$\Pi(t)$ is a random partition function that encodes the genealogy formed by the $n$ lineages. Its stochasticity follows from the stochasticity of (12) as well as the stochasticity of lineages given a single realization of (12).

3 Results

Our results characterize the lineage structure at $t = 0$ of $n$ infected cells sampled at $t = T_{\text{sample}}$. More precisely, Theorems 3.1 and 3.4, provide a random partition to which $\Pi(0)$ converges in the limit $E \to \infty$, $\mu \to 0$, with the limit taken so that $\mu^3E^2|\log(\mu E)|^m \to 0$ for any $m > 0$ and $\mu E \to \infty$. We refer to this limit as the small population limit, SPL. Throughout this work, whenever we take an unspecified limit, we mean the SPL. The $|\log(\mu E)|^m$ factor in the definition of the SPL is needed to make our arguments mathematically rigorous. Its inclusion, which reflects technical limitations in our proofs, is likely unnecessary. Intuitively, and in the context of biological discussions below, we think of the SPL as requiring simply $\mu^3E^2 \to 0$ and $\mu E \to \infty$. This viewpoint will be supported by our numerical results.

In experimental and theoretical HIV studies, $E$ has been estimated in the range $10^6$–$10^8$. In Chun et al. (1997), the number of activated CD4$^+$ cells with integrated provirus was found to average $3 \times 10^7$. Various studies have estimated that somewhere between 1 in 1000 to 1 in 80000 CD4$^+$ cells are productively infected during HIV infection, see p. 91 in Levy (1998) and references therein. Using a base of $10^{11}$ infectable lymphocytes (Levy 1998), this gives a range of approximately
The Coalescence of Intrahost HIV Lineages Under Symmetric CTL Attack

Fig. 4 Dynamics of (12) for a full escape graph with $e = 5$, $\Delta k = 0.1$, $\gamma = 3$, $\mu = 10^{-5}$, $E = 10^7$ (Color figure online)

$10^6$–$10^8$ for $E$. Presumably, $E$ varies depending on the individual and stage of infection. Mutation rates for HIV per base pair per infection event have been estimated at approximately $10^{-5}$ (Crandall 1999). Through numerical experiments, we show that Theorems 3.1 and 3.4, exact in the SPL, are a good approximation for the lineage structure formed under (12) in the parameter regime $\mu = 10^{-5}$, $E = 10^6$. In contrast, we show that the parameter regime $\mu = 10^{-5}$, $E = 10^8$ is not well approximated by the SPL. The regime $\mu = 10^{-5}$, $E = 10^7$ is a middle ground in which the SPL is a reasonable approximation, but significant error does exist. Therefore, we think of the SPL as being a limiting version of (12) when HIV has a relatively small infected cell population size.

Theorems 3.1 and 3.4 do not specify the structure of $\Pi(t)$ at times other than $t = 0$. However, the arguments we use to justify these theorems do provide some results in this direction which we mention in the Discussion section. Similarly, while our results focus on lineage structure, we make some observations regarding the stochastic dynamics of (12) in the Discussion section. As we mentioned in Sect. 2.2, the $E \to \infty$ limit does not eliminate the stochasticity of (12) in certain parameter regimes for $\mu$, and the SPL is one such a regime.

In Sect. 3.1, we present our SPL results, while in Sect. 3.2, we discuss the numerical results that connect the SPL to the parameter regimes of HIV.

3.1 Small Population Limit Results

The dynamics of (12) are composed of successive sweeps in which each variant class displaces the previous variant class as the dominant portion of the infected cell population. For example, Fig. 4 shows a realization of (12) for a full escape graph with $e = 5$, $\Delta k = 0.1$, $\gamma = 3$, $g = 0.1$, $\mu = 10^{-5}$, $E = 10^7$. The figure was generated by solving (12) numerically. Notice that stochasticity plays an important role in the dynamics of (12). Variants within a given class have identical parameters, meaning that the different trajectories within a given class seen in Fig. 4 arise due to stochasticity.

Since initially there are no variants outside of the 0th and 1st variant classes, the variants from the $i$th variant class with $i > 1$ come from $E_{i-1} \to E_i$ mutations, i.e., a mutation $v' \to v$ with $v' \in E_{i-1}$ and $v \in E_i$. The SPL scaling forces such mutations to occur during a time interval when $E_{i-1}$ variants dominate the population and $E_{i-1}, E_i$ variants are at low frequencies. During this time interval, all variants in $E_j$ with $j < i - 2$ have been driven out of the population, or nearly so, while all variants in $E_j$ for $j > i$ have yet to arise. We refer to this time interval as the $E_{i-1}$ spawning phase.
because the rise in \( \mathcal{E}_i \) variants is being driven by \( \mathcal{E}_{i-1} \rightarrow \mathcal{E}_i \) mutations. At later times, once the \( \mathcal{E}_i \) population has reached higher frequencies, \( \mathcal{E}_{i-1} \rightarrow \mathcal{E}_i \) mutations have little impact on \( \mathcal{E}_i \) variant population dynamics, and the \( \mathcal{E}_{i-1} \) spawning phase ends. The condition \( \mu \rightarrow 0 \) in the SPL insures that a variant that is being spawned cannot simultaneously spawn another variant.

During the \( \mathcal{E}_{i-1} \) spawning phase, variants in \( \mathcal{E}_{i-1} \) are increasing in population size at approximately rate \( \Delta k \) while variants in \( \mathcal{E}_i \) are increasing at approximately rate \( 2\Delta k \). To see why, recall that the \( \mathcal{E}_{i-2} \) variants dominate the infected cell population during the \( \mathcal{E}_{i-1} \) spawning phase. Since \( \mathcal{E}_{i-1} \) and \( \mathcal{E}_i \) variants are attacked by CTLs at 1 less and 2 less epitopes than \( \mathcal{E}_{i-2} \) variants, their relative fitness is given by \( \Delta k \) and \( 2\Delta k \), respectively. These dynamics are a generalized version of the well-studied Luria–Delbrück (LD) model (see Zheng 1999 for an excellent review of LD models and results). The LD model assumes a wild type population growing at rate, say, \( a \) that produces mutant types that also grow at rate \( a \). This contrasts to the growth rates \( \Delta k \) and \( 2\Delta k \) for \( \mathcal{E}_{i-1} \) and \( \mathcal{E}_i \) variants, respectively, in the \( (i-1) \)th spawning phase. For this reason, we refer to spawning phase dynamics as obeying a generalized LD model. The dynamics of the LD model have been studied for many years, and the LD distribution, which gives the number of mutant types at a given time, is well understood.

We analyze (12) by decomposing the time considered, \([0, T_{\text{sample}}]\), into a series of time intervals \([T_i, T_{i+1}]\) for \( i = 0, 1, \ldots, e-2 \) along with intervals \([0, T_0], [T_{e-1}, T_e]\), and \([T_e, T_{\text{sample}}]\). The interval \([T_i, T_{i+1}]\) is the \( \mathcal{E}_{i+1} \) spawning phase. Intervals \([0, T_0], [T_{e-1}, T_e], [T_e, T_{\text{sample}}]\) are boundary cases that do not correspond to a spawning phase. In this way, we reduce (12) to a sequence of spawning phases.

For each vertex \( v \), we define the pop value of \( v \) as the number of \( v \) variants at the beginning of the \( v \) spawning phase. More precisely, if \( v \in \mathcal{E}_{i+1} \), then the pop value of \( v \) is \( e_v(T_i) \) because \( T_i \) is the beginning of \( v \)'s spawning phase. For the linear escape graph, we can express the distribution of the \( v \)th pop value through a simple formula that is independent of other pop values, see (30). In the case of a full escape graph, the distribution of the \( v \)th pop value is given by an iterative formula that depends on other pop values, see (51).

To construct genealogies, we trace LD dynamics back in time through the various spawning periods. We show that the coalescent events produced within each spawning period are described by the \( \text{PD}(1/2, 0) \) distribution, where \( \text{PD}(\alpha, \theta) \) corresponds to the random partition specified by a Poisson–Dirichlet distribution with parameters \( (\alpha, \theta) \), see Chap. 3 of Pitman (2002). For the linear escape graph, \( \Pi(0) \) can be characterized by a concatenation of \( \text{PD}(1/2, 0) \) distributions corresponding to the different spawning periods, and we arrive at the following result.

**Theorem 3.1** Consider (12) assuming a linear escape graph. Then letting \( \Delta \) be any partition of \( n \) lineages,

\[
\lim_{\text{SPL}} P \left( \Pi(0) = \Delta \right) = P \left( \text{PD} \left( \frac{1}{2e-1}, 0 \right) = \Delta \right). \tag{14}
\]
The Coalescence of Intrahost HIV Lineages Under Symmetric CTL Attack

The description of $\Pi(0)$ for the full escape graph is more complicated. To account for the interaction between different variants, we introduce the variables $D_v$ indexed over all $v \in E_{\text{full}}$, except that $v \in E_0$ is excluded.

**Definition 3.2** For each $v \in E_1$, we define $D_v = 1$. Then we recursively define $D_v$ as follows. Suppose the $D_v$ values are known for $v \in E_{i-1}$. Set

$$D_{\max,i-1} = \max_{v \in E_{i-1}} D_v.$$  

Then, for each $v \in E_i$, set

$$D_v = \sum_{v' \in P(v)} D_{v' \rightarrow v},$$

where

$$D_{v' \rightarrow v} = \left( \frac{D_{v'}}{D_{\max,i-1}} \right)^2 \mathcal{A}_{v' \rightarrow v},$$

where

$$A = \left( \frac{k_i}{\Delta k} \right) \frac{\pi^2}{2},$$

and where $\mathcal{A}_{v' \rightarrow v}$ is an iid sample from $S(1/2, 1, 1)$, where $S(\alpha, \beta, \gamma)$ is a stable distribution with index $\alpha$, skewness $\beta$, and scale $\gamma$ (Nolan 2011; Bertoin 1996).

The distribution of $S(1/2, 1, 1)$ is that of a Levy distribution and can alternatively be characterized as the value of a $1/2$-stable process run for unit time (Bertoin 1996). As we will show, for $v \in E_{i+1}$,

$$\exp\{T_i\} \mathbb{E}^2 \sigma^2 \to D_v,$$

meaning that $D_V$ is proportional to the pop value of $v$ with constant independent of $v$.

For the full escape graph, the state of our lineages is not simply a partition of $\{\ell_1, \ell_2, \ldots, \ell_n\}$. Rather, we must specify a vertex to which each lineage is associated at a given time $t$. Intuitively, the vertex associated with, say, $\ell_j$ at time $t$ is the variant type of the infected cell at time $t$ from which the $j$th sampled cell descends. To put this in the context of a partition function, $\Pi(t)$ partitions the lineages into disjoint sets and associates with each such set a variant in $E$. The $\mathcal{E}_i$ defined below are random partitions for which every element of the partition is associated with a variant in $E_{i+2}$. To explain the $\mathcal{E}_i$, we need to discuss $\text{PD}(1/2, 0)$ in greater detail. Let $x(t)$ be a $1/2$-stable process, and let $J_1, J_2, \ldots$ be the jumps of this process, in any particular order, that occur up to time $t$. Then $\text{PD}(1/2, 0)$ is determined by the random partition of $[0, 1]$ with intervals of size $J_i/x(1)$. (See Pitman 2002, Chap. 4 for a more thorough discussion of these issues.) We can consider this partition conditioned on a particular value of $x(1)$. We label this random partition $\text{PD}(1/2, 0 \mid x(1) = X)$ or more succinctly $\text{PD}(1/2, 0 \mid X)$. 

\[\text{Springer}\]
Definition 3.3 We define a partition $\Xi_i(L)$ on a set $L$ for which each element $\ell \in L$ is associated with a vertex $v_\ell \in E_{i+2}$. For every $v_\ell$, choose $p_\ell \in P(v_\ell)$ according to the probabilities

$$P(p_\ell = v') = \frac{D_{v' \rightarrow v}}{D_v}.$$  

(20)

Partition $L$ into sets $U_{v',v}$ for $v' \in P(v)$ and $v \in E_{i+2}$ defined by

$$U_{v',v} = \{ \ell \in L : v_\ell = v, p_\ell = v' \}.$$  

(21)

Let $U^{PD}_{v',v}$ be a random partition sampled from the PD($1/2, 0 | S_{v' \rightarrow v}$) distribution applied to the elements of $U_{v',v'}$. Associate with each element of $U^{PD}_{v',v}$ the variant $v'$. Then we define $\Xi_i(L)$ as the partition of $L$ formed by $\bigcup U^{PD}_{v',v}$.

Intuitively, the $U_{v',v}$ group lineages according to their associated variant type when a spawning period ends, $v$, and when the spawning period begins, $v'$. Within each such grouping, coalescent events occur according to PD($1/2, 0 | S_{v' \rightarrow v}$).

Theorem 3.4 Consider (12) assuming a full escape graph. Then letting $\Delta$ be any partition of the $n$ samples,

$$\lim_{\text{SPL}} P(\Pi(0) = \Delta) = P\left(\left(\prod_{j=0}^{e-2} \Xi_j\right) = \Delta\right),$$

(22)

where $\Xi_i$ is given by Definition 3.3

For the full escape graph, each of the $n$ lineages is initially (at time $T_{\text{sample}}$) associated with a 11...1 variant. By $\prod_{j=0}^{e-2} \Xi_j$ we mean that first $\Xi_{e-2}$ is applied to the $n$ uncoalesced lineages at $T_{\text{sample}}$, then $\Xi_{e-3}$ is applied to partition resulting from $\Xi_{e-2}$, and so on recursively. In the full escape graph, we associate each lineage at any given time with a variant. In particular, this is true at $t = 0$, and correspondingly $\Delta$ should assign to each lineage a variant in $\hat{E}_{\text{full}}$. However, for simplicity, Theorem 3.4 refers only to the partition structure of the lineages at $t = 0$ and ignores the labeling.

3.2 Numerical Results

In this section, we consider five parameter regimes: small population limit (SPL), the approximating regime (AR), the small population regime (SPR), the medium population regime (MPR), and the large population regime (LPR). Table 2 specifies the $\mu$ and $E$ value associated with each regime. The table also includes the corresponding values for $\mu^3 E^2$ and $\mu E$. The AR has $\mu^3 E^2 < 1$ while $\mu E \gg 1$, suggesting a good approximation by the SPL. Notice that the SPR has a scaling near the SPL, but that the LPR has a $\mu^3 E^2$ value of 10 which, as we shall show, is too large for the SPL to apply.

To understand the accuracy of the SPL, we first consider the probability that two sampled lineages coalesce. More precisely, setting $n = 2$, we consider the probability that $\ell_1$ and $\ell_2$ coalesce by $t = 0$ or, equivalently, $P(\Pi(0) = \{[\ell_1, \ell_2]\})$. This
probability is often computed in population genetics applications and is one way to characterize $\Pi(0)$ (Wakeley 2008). Figure 5 shows this coalescent probability for a linear escape graph with $\gamma = 3$, $g = 0.1$, $\Delta k = 0.1$. The bars, from left to right, give the coalescent probability under the SPL, AR, SPR, MPR, and LPR (see Table 2 for the definition of these parameter regimes).

The next four bars represent, from left to right, the coalescent probability for the AR, SPR, MPR and LPR, respectively. These values are computed by solving (12) numerically and forming lineages on top of the stochastic dynamics. We compute 1000 realizations of (12) dynamics, and for each such realization, we consider the coalescence of two lineages 1000 times. We then average over all 1000 lineage pairs and all 1000 realizations. Solving (12) and building lineages on top of the dynamics is not numerically trivial due to the large population size. Following methods described in Leviyang (2012), we solve (12) exactly and track parent–child relationships in each variant until the variant population size exceeds 10000. At that point we switch to the deterministic ODE analogue of (12).
The probability of coalescence of two lineages for a full escape graph with $\gamma = 3$, $g = 0.1$, and $\Delta k = 0.1$. The bars, from left to right, give the coalescent probability under the SPL, AR, SPR, MPR, and LPR (see Table 2 for the definition of these parameter regimes).

As Fig. 5 demonstrates, the SPL is a good approximation in the AR and SPR, but not the LPR. The MPR represents a middle ground. Figure 6 is the same as Fig. 5, except that in Fig. 6, we consider a full escape graph. In this case, a simple formula such as (23) is not possible. Instead, we sample the $D_{v'}, S_{v'} \rightarrow v$ as defined in Definition 3.2 and construct partitions according to Definition 3.3 and Theorem 3.4. Explicit formulas for the distribution of $P\left(D_{v'}, S_{v'} \rightarrow v \mid X\right)$ are available, see, for example, Pitman (2002), Chap. 4, Lemma 4.2. For a large number of samples, the formulas become extremely complex, but in the case of two samples, the probability of coalescence is given by a relatively simple expression,

$$P \left(D_{v'}, S_{v'} \rightarrow v \mid X\right) = \left\{ \ell_1, \ell_2 \right\} = 1 - \frac{\exp \left( \frac{1}{2t} \right)}{\sqrt{t}} \int_{-\infty}^{\infty} dx \exp \left[ - \frac{z^2}{2t} \right].$$

We sample the $D_{v'}, S_{v'} \rightarrow v$ 1000 times, and for each such sampling, we sample the $\Xi_i$ partitions 1000 times by exploiting (24).

Another value that can be used to characterize the HIV genealogy shaped by CTL attack is the number of still uncoalesced lineages at $t = 0$. That is, the number of elements in $\Pi(0)$. Recall that each element of $\Pi(0)$ is a collection of lineages that have coalesced. Figure 7 shows the distribution of this number for a full escape graph with $e = 3$, $\gamma = 10$, $\Delta k = 0.3$, and $n = 100$. The same pattern of accuracy is seen as with Figs. 5 and 6. As with Fig. 6, results were generated through simulation. However, in this case, the formulas for $P\left(D(1/2, 0 \mid S_{v'} \rightarrow v)\right)$ are too complex to implement. Instead, we simulate the birth–death process $I(t)$ defined in Proposition 5.4 keeping track of birth and death events. The simulation provides a value of $S_{v'} \rightarrow v$, and by tracking birth and death events, we are able to repeatedly sample lineages associated with that given $S_{v'} \rightarrow v$. Alternatively, we could simulate a $1/2$-stable process conditioned on its end state using standard methods (Asmussen and Glynn 2007), but we did not use this approach.

Theorems 3.1 and 3.4 provide a theoretical framework for understanding genealogies on (12). However, they also provide a computational approach for sampling such genealogies that is much faster than solving (12) directly. Table 3 gives the CPU time in seconds required to generate the coalescent probability results shown in Figs. 5...
The number of uncoalesced lineages at $t = 0$ assuming a sample of 100 infected cells after HIV escape. Results correspond to a full escape graph with $e = 3$, $\gamma = 10$, and $\Delta k = 0.3$. The bars, from left to right, give values under the SPL, AR, SPR, MPR, and LPR (see Table 2 for the definition of these parameter regimes).

Table 3  CPU time in seconds needed to generate coalescent probabilities of Figs. 5 and 6. CPU is an Intel i7-2600

| Graph     | $e$ | SPL time | SPR time | SPR/SPL |
|-----------|-----|----------|----------|---------|
| LINEAR    | 2   | 3        | 1900     | 630     |
|           | 6   | 13       | 6300     | 480     |
| FULL      | 2   | 4        | 2400     | 600     |
|           | 6   | 27       | 30400    | 1130    |

and 6 for the cases $e = 2, 6$. We show the CPU times needed to produce the probability through our SPL results, and by solving (12) in the SPR, the times required for the APR, MPR, and LPR are similar to the SPR. As can be seen, the SPL approach provides significant computational savings.

4 Discussion

Application of the results we have presented depends on approximating the SPL scaling by satisfying $\mu E \gg 1$ and $\mu^3 E^2 \ll 1$. HIV almost certainly always satisfies $\mu E \gg 1$, so this condition is not restrictive. On the other hand, $\mu^3 E^2 \ll 1$ is satisfied if the HIV population size is of relatively small magnitude, namely on the order of $10^6$. Importantly, the HIV population size we must consider is the number of productively infected CD4$^+$ cells. Inactivated CD4$^+$ cells or those infected by nonfunctional HIV do not enter into our model because they do not produce offspring infected cells.

Our results have implications for both dynamics and genealogies. For dynamics, our arguments show that the stochasticity of (12) in the SPL is completely contained within the pop values described in the results section and defined precisely in (30) and (51). Intuitively, once a variant population reaches large size, averaging effects take over and deterministic dynamics apply. In our nomenclature, a variant population is small and hence experiences stochastic dynamics only when it is being spawned by another variant population. If we are interested in dynamics and not lineages, then (12) can be reduced to a deterministic ODE accompanied by stochastic pop values.
The stochasticity of the pop values has significant impact on the dynamics of (12). The left graph of Fig. 8 gives a solution for the deterministic analogue of (12) in which stochastic events are replaced by their average. Another way to describe such a system is as (12) when all pop values are equal. Either way, since our equations are symmetric, the dynamics must be symmetric, and this is indeed the case in left graph. All variants within the same variant class have identical dynamics.

In contrast, the right graph of Fig. 8 provides the dynamics for a single realization of (12). We can see that stochasticity plays an essential role in (12) because the left and right graphs represent very different dynamics. Our work explains the stochasticity seen in the right graph. The variant 011 dominates 110, 101 because of a relatively high pop value, i.e., due to stochastic effects. Biologically, the stochasticity of pop values come from the stochasticity of mutation times. Critically, this stochasticity depends not on the existence of few infected cells, but rather on the relationship between mutation rates and population sizes. For example, the AR regime (see Table 2) assumes a population size of \(O(10^{13})\), yet the dynamics of (12) are still stochastic. Similarly, the MPR regime, which based on current experimental results is likely to be a regime appropriate to HIV, shows similar stochastic behavior with a population size \(O(10^7)\).

Turning now to genealogies, we have described the coalescence of lineages caused by the whole period of HIV escape. However, as mentioned, we can decompose HIV escape into time intervals \([T_i, T_{i+1}]\). Each such period corresponds to a PD(1/2, 0) or \(\Xi_i\) partition, and so we know the distribution of the lineages at each \(T_i\) given the lineage state at \(T_{i+1}\). Between the \(T_i\), however, our results do not describe the lineages. Figures 9 and 10 show genealogies formed for a 5 epitope and 2 epitope attack, respectively, in the case of a linear escape graph under the SPR. Here we have shown all coalescent events that happen during \([T_i, T_{i+1}]\) to occur at \(T_i\). Both genealogy figures were produced using Figtree. (Figtree is available as part of the BEAST software package (Drummond and Rambaut 2007).)

Our results have implications for intrahost HIV effective population size, a topic explored by Kouyos et al. (2006), Leigh-Brown (1997), Rouzine and Coffin (1999). Effective population size, as opposed to census population size, is a measure of the potential genetic diversity of a population. HIV experimental results show a relatively
small level of genetic diversity and hence a small effective population size (Leigh-Brown 1997). In this work, we have assumed a large census size given by $E$. In contrast, we see through Figs. 5 and 6 that CTL attack at several epitopes causes coalescence of lineage pairs with high probability, a fact that leads to small effective population size. To quantify the effect of CTL attack on effective population size, we would need a model for repeated attacks that occur over the course of infection. We have not pursued this issue, but our current results suggest an important role for CTL attack in shaping HIV effective population size.

The restriction of our current results to symmetric attack and the small end of the HIV population size range is a significant limitation. Further work should allow for these restriction to be lifted, but our current results provide some general observations.

For a full escape graph, the assumptions of symmetric attack makes the paths through the graph identical in terms of the underlying parameters. Removal of the
symmetric attack assumption would lead to a dominant path. For example, if there is an epitope that is attacked more strongly than the other epitopes, then it will be the first epitope at which HIV escapes CTL attack. Of course, there will be some HIV variants that initially posses a mutation at a different epitope, but these will be few in number. The order of the epitopes at which HIV escapes from the CTL attack will be specified in the case of asymmetric CTL attack. As a result, HIV escape on a full escape graph in the asymmetric attack case should proceed essentially on one path of the graph and be similar to the linear escape graph dynamics and genealogies we have discussed.

Figure 11 compares coalescent probabilities for linear and full escape graphs. This is the same data presented in Figs. 5 and 6. In Fig. 11, the four bars give, from left to right, the coalescent probability for a linear escape graph under SPR, a full escape graph under SPR, a linear escape graph under LPR, and a full escape graph under LPR. As can be seen, the coalescent probabilities under the SPR are similar for the linear and full escape graphs. Some numerical experiments suggest that this is because pop value stochasticity causes a single path through the full escape graph to dominate, similarly to our comments in the previous paragraph on asymmetric attack, but we have not pursued this issue.

5 Linear Escape Graph

In this section, we consider (12) for the linear escape graph under the SPL. Our main aim is to explain and demonstrate Theorem 3.1. For notational simplicity, we set $v_i = \underbrace{11 \ldots 100 \ldots 0}_i$ and write $e_i$ for $e_{v_i}$. For each variant class $\mathcal{E}_i$, we define $T_i$ for $i = 1, 2, \ldots, e$ as the time at which variant $v_i$ reaches scaled population size $\delta$,

$$T_i = \inf\{t : e_i \geq \delta\},$$

where

$$\delta = \left(\frac{1}{|\log(\mu^2E)|}\right)^2.$$

(26)
The Coalescence of Intrahost HIV Lineages Under Symmetric CTL Attack

Table 4 Dynamics of (12) during $[T_i, T_{i+1}]$ for a linear escape graph

| Variant | $e_i(T_i)$ | $e_i(T_{i+1})$ |
|---------|------------|----------------|
| $v_{i-1}$ | $(1 - h(T_i))/h(T_i) + O(\delta)$ | $O(\delta)$ |
| $v_i$ | $\delta$ | $(1 - h(T_{i+1}))/h(T_{i+1}) + O(\delta)$ |
| $v_{i+1}$ | $O(\mu^2\delta^2)$ | $\delta$ |
| $v_{i+2}$ | $0$ | $O(\mu^2\delta^2)$ |
| $v_j$ for $j > i + 2$ | $0$ | $0$ |
| $v_j$ for $j < i - 1$ | $o(\delta)$ | $o(\delta)$ |

The value of $\delta$ can fall within a range of values, the formula above is a specific choice within this range. Different variants “interact” in (12) through the $h$ equation. When a variant has population less than $\delta$, its impact on $h$ dynamics and in turn on the dynamics of other variants is small and can be ignored in the SPL. From this perspective, the smaller $\delta$, the better. On the other hand, a $\delta$ that is too small will make the interval $[T_{i-1}, T_i]$ too short in the sense that the $v_i \rightarrow v_{i+1}$ mutations that drive the $i$th spawning period will not have finished by $T_i$. From this perspective, the larger $\delta$, the better. Our choice for $\delta$ is a middle ground between these two extremes.

In the SPL, $\delta \rightarrow 0$. Variants with scaled population size less than $\delta$ collapse as a percentage of the population in the SPL. However, if a variant has scaled population size $\delta$, then the number of such variants, unscaled, is $\delta \mathbb{E}$ which goes to $\infty$ in the SPL.

We also set

$$T_0 = \inf \{ t : e_1(t) = \mu^2 \mathbb{E} \delta^2 \}.$$  \hspace{1cm} (27)

$T_0$ is a special case because we set $e_1(0) = \mu \mathbb{E}$.

The $T_i$ decompose $[0, T_{\text{sample}}]$ into intervals $[T_{i-1}, T_i]$ along with initial and final intervals $[0, T_0]$ and $[T_e, T_{\text{sample}}]$, respectively. That this, composition is valid with probability 1 in the SPL, i.e.,

$$P(T_0 < T_1 < \cdots < T_e < T_{\text{sample}}) \rightarrow 1$$ \hspace{1cm} (28)

will be a consequence of our analysis below.

We consider the interval $[T_i, T_{i+1}]$ for $i = 0, 1, \ldots, e - 2$. The intervals $[0, T_0]$, $[T_{e-1}, T_e]$, $[T_e, T_{\text{sample}}]$ are handled separately. During $[T_i, T_{i+1}]$, only the variants $v_{i-1}, v_i, v_{i+1},$ and $v_{i+2}$ play a significant role in the dynamics. Table 4 shows the scaled population sizes of different variants at $T_i$ and $T_{i+1}$. The arguments that justify Table 4 are given below, for now we focus on intuition.

To explain Table 4, we first consider the $v_{i-1}$ and $v_i$ variants. If only one variant type exists in the population, say $v$, then (12) is composed solely of the equations for $h$ and $e_v$, and in equilibrium we have $e_v \approx (1 - h)/h$. Examining Table 4, we see that at $T_i$, $v_{i-1}$ is roughly at this equilibrium, meaning that it is the dominant variant in the HIV population. On the other hand, $v_i$ variants at time $T_i$ are few since $\delta \ll (1 - h)/h$. However, by time $T_{i+1}$, the situation has flipped with $v_i$ dominating the population and $v_{i-1}$ pushed to low levels. Intuitively, the $v_i$ variants are more fit and push out the $v_{i-1}$ variants during $[T_i, T_{i+1}]$. 

\[ \text{Springer} \]
Now consider the \( v_{i+1}, v_{i+2} \) variants. Recalling that \( \mathbb{E}e_{i+1}(T_i) \) gives the number, unscaled, of \( v_{i+1} \) variants, note that

\[
\mathbb{E}e_{i+1}(T_i) = O\left( \mu^2 \mathbb{E}^2 \delta^2 \right) \ll \frac{1}{\mu}.
\]  
(29)

The \( \ll \) directly above is justified in the SPL since \( \mu^3 \mathbb{E} \to 0 \). By the same observation, the rate of \( v_{i+1} \to v_{i+2} \) mutations at \( T_i \), \( \mu \mathbb{E}e_{i+1}(T_i) \) goes to 0 in the SPL. However, notice that \( \mathbb{E}e_{i+1}(T_i) \approx O(\mu^2 \mathbb{E}^2) \to \infty \), meaning that \( e_{i+1} \) dynamics are deterministic at time \( T_i \) in the SPL.

Turning to \( v_{i+2} \), we see that at \( T_i \) no such variants exist. However, by time \( T_{i+1} \), there are enough such variants to make their dynamics deterministic. Connecting to our comments in the Results section, \( [T_i, T_{i+1}] \) is the \((i+1)\)th spawning phase or, slightly more explicitly, the \( v_{i+1} \to v_{i+2} \) spawning phase.

Finally, we note that Table 4 shows that all other variant types are of negligible population size. \( v_j \) with \( j > i + 2 \) have yet to arise, and \( v_j \) with \( j < i - 1 \) have been previously driven to low levels by fitter variants.

Table 4 provides the outlines of an iteration, as we proceed through different values of \( i \), that describes the stochastic dynamics of (12). The key to deriving the table is an estimate of \( e_{i+2}(T_{i+1}) \), the pop value of \( v_{i+2} \). In Sect. 5.1, we show

\[
\frac{e_{i+2}(T_{i+1})}{\mu^2 \mathbb{E} \delta^2} \to A \mathcal{S},
\]  
(30)

where \( A \) and \( \mathcal{S} \) are as defined in Definition 3.2. In Sect. 5.1, we provide the arguments that justify Table 4 and (30). Then in Sect. 5.2, we use the results of Sect. 5.1, which center on the dynamics of (12), to demonstrate our lineage result, Theorem 3.1.

Unless otherwise stated, all limit relations involving r.v.’s are meant in the sense of distribution. For a r.v. \( X \) and \( f(\mu, \mathbb{E}) \in \mathbb{R} \), we write \( X = O_p(f(\mu, \mathbb{E})) \) if for all \( \epsilon > 0 \), there exists \( K \) such that

\[
\limsup_{\text{SPL}} P\left( X > K f(\mu, \mathbb{E}) \right) < \epsilon,
\]  
(31)

and we write \( X = o_p(r) \) if

\[
\lim_{\text{SPL}} X/f(\mu, \mathbb{E}) = 0.
\]  
(32)

5.1 Dynamics

The goal of this section is to prove Proposition 5.1 which is a precise version of Table 4 and (30). Throughout this section and below, the stochastic dynamics of a variant with scaled population size exceeding \( O(\mu^2 \mathbb{E} \delta^2) \) will be replaced by corresponding deterministic dynamics. More precisely, during \([T_i, T_{i+1}]\), all variants \( e_j \) with \( j \leq i \) will be taken to evolve deterministically, and \( e_{i+1} \) will be taken to evolve deterministically on \([T_i^h, T_{i+1}]\) where \( T_i^h \) will be defined below.

If for some variant \( v, e_v > \mu^2 \mathbb{E} \delta^2 \), then the number of \( v \) infected cells is greater than \( \mu^2 \mathbb{E}^2 \delta^2 \). Under the SPL, \( \mu^2 \mathbb{E}^2 \delta^2 \to \infty \). Heuristically, if a variant’s population...
size goes to infinity in the SPL, then a law of large numbers will force the stochastic dynamics to converge to deterministic dynamics. We make use of this heuristic throughout this section in order to simplify the arguments and provide intuition. A rigorous justification of this approach is given in Appendix A.4.

**Proposition 5.1** Consider (12) on a linear escape graph. Assume that at $T_i$ for $i = 1, \ldots, e - 2$:

1. $P(e_j(T_i) = 0) \to 1$ for $j \geq i + 2$.
2. $\frac{e_{i+1}(T_i)}{\mu^2 \mathbb{E}^2} \to A^*.$
3. $e_i(T_i) = \delta.$
4. $e_{i-1}(T_i) - \frac{1 - h(T_i)}{h(T_i)} = O_p(\delta).$
5. $e_j(T_i) = o_p(\delta)$ for $j < i - 1$.

Then at time $T_{i+1}$ we have the following conclusions:

1. $P(e_j(T_{i+1}) = 0) \to 1$ for $j \geq i + 3$.
2. (30) holds.
3. $e_{i+1}(T_{i+1}) = \delta.$
4. $e_i(T_{i+1}) - \frac{1 - h(T_{i+1})}{h(T_{i+1})} = O_p(\delta).$
5. $e_j(T_{i+1}) = o_p(\delta)$ for $j < i$.

Conclusion 3 of Proposition 5.1 holds by the definition of $T_{i+1}$ as long as $\delta \mathbb{E}$ is an integer. We make this simplifying assumption to avoid technicalities.

Proposition 5.1 can be applied recursively to characterize the dynamics at each time $T_i$ for $i = 1, \ldots, e - 2$. The case $i = 0$, which we must consider to start the recursion, is handled through the same arguments that give Proposition 5.1, except that our assumptions are slightly different. Namely, at $T_0$ we have $e_1(T_0) = \mu^2 \mathbb{E}^2$ by the definition of $T_0$, which parallels assumption 2 of Proposition 5.1, and $e_0(T_0) = (1 - h)/h + O_p(\delta)$, which parallels assumption 4. Assumption 1 of Proposition 5.1 holds for the $i = 0$ case, while assumptions 3 and 5 are not applicable.

To demonstrate Proposition 5.1, we split $[T_i, T_{i+1}]$ into two time intervals, $[T_i, T_i^h]$ and $[T_i^h, T_{i+1}]$. In Lemma 5.2, we show that during $[T_i, T_i^h]$, the $v_i$ variant displaces the $v_{i-1}$ variant as the dominant variant, as alluded to in Table 4 and the accompanying discussion. This transition happens quickly, so that $T_i^h - T_i$ is small. As a result, the $v_{i+1}$ population does not grow in size much, and no $v_i \to v_{i+2}$ mutations occur. Through the arguments of Lemma 5.3, we show that $v_{i+1}, v_{i+2}$ dynamics on $[T_i^h, T_{i+1}]$ obey the generalized LD dynamics discussed in the Results section.

**Lemma 5.2** Adopt the same assumptions stated in Proposition 5.1. Set $T_i^h = T_i + (3/\Delta k + \frac{2}{\rho} \log(\delta))$, where $\rho$ is given in (66). Then,

1. for $t \in [T_i, T_i^h]$, $\mu \mathbb{E} e_{i+1}(t) \to 0$,
2. $P(e_{i+2}(T_i^h) = 0) \to 1$,
3. \( e_i(T^h_i) - \frac{1 - h(T^h_i)}{h(T^h_i)} = O_p(\delta) \),
4. \( e_{i-1}(T^h_i) = O_p(\delta) \).

Conclusions 1 and 2 of Lemma 5.2 guarantee that no \( v_{i+1} \rightarrow v_{i+2} \) mutations occur during \([T_i, T^h_i]\) and follow from the small population size of \( v_{i+1} \) variants at \( T_i \). Indeed, by assumption 2, we can show that for any \( \epsilon > 0 \), there exist \( c_1, c_2 > 0 \) such that
\[
\lim_{SPL} P(c_1 \mu^2 \delta^2 < e_{i+1}(T_i) < c_2 \mu^2 \delta^2) > 1 - \epsilon. \tag{33}
\]
Referring to (12), we note that \( h \) is bounded above by 1, and so we can bound the expected growth rate of \( e_{i+1} \) by \( \gamma \). Standard branching process results then show that \( \exp[-\gamma(t - T_i)]e_{i+1}(t)/e_{i+1}(T_i) = O_p(1) \) (Athreya and Ney 1972). Using the assumption on \( e_{i+1}(T_i) \) and the fact \( T^h_i - T_i = O(\log(\delta)) \), we find for \( t \in [T_i, T^h_i] \),
\[
\mu \mathbb{E} e_{i+1}(t) = O_p(\mu^3 \delta^2 - c) \rightarrow 0, \tag{34}
\]
where \( c = \gamma(3/\Delta k + 2/\rho) \). This gives Conclusion 1.

Conclusion 2 follows almost directly from Conclusion 1. We recall from (12) that \( v_{i+1} \rightarrow v_{i+2} \) mutations arise at rate \( \mu \mathbb{E} \gamma h e_{i+1} \). The number of such mutations in the interval \([T_i, T^h_i]\) is then a Poisson distribution with mean
\[
O(\mu \mathbb{E}) \int_{T_i}^{T^h_i} ds h(s)e_{i+1}(s) \tag{35}
\]
Plugging into (34) shows that the mean number of mutations goes to 0 under the SPL. Applying a Chebyshev argument gives Conclusion 2.

To explain Conclusions 3 and 4 of Lemma 5.2, we notice that only variants \( v_i \) and \( v_{i-1} \) are of order greater than \( \delta \) throughout \([T_i, T^h_i]\). Ignoring the other variants then, (12) reduces to three equations involving \( e_i, e_{i-1}, h \). Since variant \( v_i \) has one less epitope exposed to CTL attack than \( v_{i-1} \), it will eventually push the \( v_{i-1} \) to extinction. Initially, \( e_i(T_i) = \delta \). Since the CTL kill rate of \( v_{i-1} \) variants is \( \Delta k \) greater than those of \( v_i \) variants, initially the \( v_i \) variants grow exponentially with rate \( \Delta k + O(\delta) \). It then takes \( O(\frac{1}{\Delta k} \log(\delta)) \) time for the \( v_i \) population to rise to \( O(1) \) levels, push out the \( v_{i-1} \) population, and near equilibrium with respect to \( h \). This explains the order of \( T^h_i \). The exact form of \( T^h_i \) is explained in Sect. A.1 of the appendix as are the technical details demonstrating Conclusions 3 and 4.

Now we consider \([T^h_i, T_{i+1}]\) through the following lemma.

**Lemma 5.3** Assume the conclusions of Lemma 5.2. Then,
1. \( e_{i+1}(T_{i+1}) = \delta \),
2. \( \frac{e_{i+2}(T_{i+1})}{\mu^2 \delta^2} \rightarrow A\mathbb{S} \),
3. \( e_i(T_{i+1}) - \frac{1 - h(T_{i+1})}{h(T_{i+1})} = O_p(\delta) \),
4. \( e_j(T_{i+1}) = o_p(\delta) \) for \( j < i \).
Conclusion 1 of Lemma 5.3 follows from the definition of $T_{i+1}$. Assuming Conclusion 1, we see that variants $v_{i+1}$, $v_{i+2}$ remain at $O_{p}(\delta)$ levels throughout $[T^{h}, T_{i+1}]$. As a result, the approximate equilibrium of $e_{i}$, $h$ which exists at $T^{h}$ is maintained. Further, variants $v_{j}$ for $j < i$ continue to drop in number as they are less fit than $v_{i}$ variants. Putting all these comments together justifies Conclusions 3 and 4.

We have left to consider Conclusion 2. From (12) we have the following ODEs for $e_{i+1}, e_{i+2}$:

\[ de_{i+1} = \gamma e_{i+1} \left( h - \frac{k_{i+1}}{\gamma} \right), \]
\[ de_{i+2} = \frac{1}{\gamma} \left( dP(h E e_{i+2}) - dP(k_{i+2} E e_{i+2}) + dP(\mu \gamma E e_{i+1}) \right). \]

(36)

By Conclusion 3, since $e_{i}$, $h$ are near equilibrium, we know $h = k_{i}/\gamma + O(\delta)$. Plugging this result into (36) gives

\[ de_{i+1} = (\Delta k + O(\delta)) e_{i+1}, \]
\[ de_{i+2} = \frac{1}{\gamma} \left( dP\left((k_{i} + O(\delta)) E e_{i+2}\right) - dP(k_{i+2} E e_{i+2}) + dP(\mu(k_{i} + O(\delta)) E e_{i+1}) \right). \]

(37)

If we label the number of $v_{i+2}$ variants as $e^\#_{i+2}$, by our scaling $e^\#_{i+2} = E e_{i+2}$, the $e_{i+2}$ equation in (37) transforms into

\[ de^\#_{i+2} = dP\left((k_{i} + O(\delta)) e^\#_{i+2}\right) - dP(k_{i+2} e^\#_{i+2}) + dP(\mu(k_{i} + O(\delta)) E e_{i+1}). \]

(38)

and we find that $v_{i+2}$ variants evolve according to a binary branching process with birth rate $k_{i} + O(\delta)$, death rate $k_{i+2}$, and mutation rate that creates new $v_{i+2}$ variants $\mu(k_{i} + O(\delta)) E e_{i+1}$.

By the assumptions of Lemma 5.3, there are no $v_{i+2}$ variants at time $T^{h}_{i}$. The $v_{i+2}$ population arises from mutations in the $v_{i+1}$ population which expands at rate $\Delta k$. Such mutations produce $v_{i+2}$ cells that then expand at expected rate $2 \Delta k$, precisely the generalized LD dynamics mentioned in Sect. 3.

We define $LD_{\text{classic}}(t)$ to be the number of mutants at time $t$ for the LD model in which mutants and wild types grow at the same rate. In Kepler and Oprea (2001), Mohle (2005), the following asymptotic formula was derived under the further assumptions that wild types grow deterministically and mutants grow stochastically, but with no death events:

\[ \lim_{m \to \infty} \frac{LD_{\text{classic}}(t) - \log(m)}{m} = S(1, 1, \frac{\pi}{2}), \]

(39)

where $m$ is the expected number of mutations on the time interval $[0, t]$ for a wild type population that is of size 1 at time 0 (the meaning of $S(\cdot)$ was specified in Definition 3.2). The relative error of (39) goes to 0 as $m \to \infty$. Typically $m \to \infty$ by keeping all parameters constant and taking $t \to \infty$.

The following proposition generalizes (39) to general growth rates. Its proof is given in Appendix A.2.
Proposition 5.4 Let $I(t)$ be the number of individuals in a population at time $t$. Assume that immigrants arrive at rate $f(t) = \mu \exp[r_1 t]$ with $\mu, r_1 > 0$ and that individuals at time $t$ are composed of descendants from immigrants on $(-\infty, t]$. Individuals in the population produce offspring independently according to a continuous-time binary branching process with birth and death rates $b$ and $d$, respectively. Let $r_2 = b - d$ and assume that $r_2 > r_1$. For a given time $T$, set

$$\omega = \left(\mu \exp[r_1 T]\right)^{\frac{r_2}{r_1}}.$$  

(40)

Then fixing $r_1, r_2$, but letting $\omega \to \infty$ through any scaling of $\mu$ and $T$, we have

$$\lim_{\omega \to \infty} \frac{I(T)}{\omega} = A_{\text{gen}}S\left(\frac{r_1}{r_2}, 1, 1\right),$$  

(41)

where

$$A_{\text{gen}} = \left(\frac{1}{b} \left(\frac{b}{r_2}\right)^{\frac{r_1}{r_2}} \frac{\Gamma(1 + \frac{r_1}{r_2}) \Gamma(1 - \frac{r_1}{r_2})}{\Gamma\left(1 + \frac{r_1}{r_2}\right)}\right)^{\frac{r_2}{r_1}}.$$  

(42)

The $A$ in Definition 3.2 is $A_{\text{gen}}$ made specific to the case of spawning dynamics. In Appendix A.3, we apply the proposition to the case of a spawning interval, $[T_i, T_{i+1}]$, to give Conclusion 2 of Lemma 5.3.

5.2 Lineage Construction

To demonstrate Theorem 3.1, we need some additional lineage notation. Recall that we consider $n$ lineages, $\ell_j$ for $j = 1, 2, \ldots, n$, corresponding to $n$ sampled cells. We let $\ell_j(t)$ be the ancestral cell at time $t$ of sample cell $j$. (To make this precise, we could number the cells in our process as they are born, and then $\ell_j(t)$ would map to $N$ but we will not make this explicit.) We let $V(\ell_j(t))$ be the variant type of $\ell_j(t)$. For example, $V(\ell(T_{\text{sample}})) = 11 \ldots 1$. We write $\ell_j$, dropping the time dependence, when we are considering the lineage over a range of times.

To combine the separate lineages into a genealogy, we need to identity mutation and coalescent events. A mutation event on $\ell_j$ occurs at time $t$ if the variant of $\ell_j$ changes at time $t$, more precisely, $V(\ell_j(t-)) \neq V(\ell_j(t))$. Given two lineages $\ell_j, \ell_k$, the lineages have coalesced by time $t$ if $\ell_j(t) = \ell_k(t)$, and we say that the lineages coalesced at time $t$ if for $t' > t$, $\ell_j(t') \neq \ell_k(t')$.

We prove Theorem 3.1 by considering mutation and coalescent events on the interval $[T_i, T_{i+1}]$. Lemma 5.7 sets up this analysis by showing that no mutation or coalescent events occur on $[T_{e-1}, T_{\text{sample}}]$. This allows us to consider the interval $[T_{e-2}, T_{e-1}]$ with all lineages of type $v_e$ at $T_{e-1}$.

Lemma 5.8 provides two results for the general setting of an interval $[T_i, T_{i+1}]$, assuming that all lineages are of type $v_i+2$ at $T_{i+1}$. First, by time $T_i$, all lineages are of type $v_i+1$. This result implies that each lineage must experience a $v_i+1 \to v_i+2$ mutation during $[T_i, T_{i+1}]$. Second, two lineages, $\ell_j, \ell_k$ coalesce during $[T_i, T_{i+1}]$ if and only if their associated cells at $T_{i+1}$, $\ell_j(T_{i+1}), \ell_k(T_{i+1})$ descend from the same $v_i+1 \to v_i+2$ mutation. Lemma 5.8 reduces the analysis of coalescent events on
The comments of the previous two paragraphs allow us to build lineages on $[T_i, T_{i+1}]$ by tracking $v_{i+1} \rightarrow v_{i+2}$ mutations and the associated number of descendants at $T_{i+1}$. In a general setting, the dynamics of the spawning period are given by Proposition 5.4, and we have the following proposition that describes lineages. The proof is given in Appendix A.2 and depends on results from the Poisson–Dirichlet literature (Pitman 2002; Pitman and Yor 1997; Perman et al. 1992).

**Proposition 5.5** Consider the assumptions of Proposition 5.4. Assume that at time $T$ we sample $n$ individuals uniformly from the population and let $\Pi$ be the partition which groups individuals descending from the same immigrant. Then,

$$\lim_{\omega \to \infty} \Pi = \text{PD}\left(\frac{r_1}{r_2}, 0\right).$$

In Appendix A.3, we use this proposition to demonstrate the following lemma, which is specific to spawning dynamics.

**Lemma 5.6** Let $\ell_1, \ell_2, \ldots, \ell_n$ be lineages corresponding to $n$ variants of type $v_{i+2}$ sampled at time $T_{i+1}$. Let $\Pi$ be the partition of these lineages at time $T_i$. Then,

$$\lim_{\text{SPL}} \Pi = \text{PD}\left(\frac{1}{2}, 0\right).$$

Existing results show that the concatenation of a PD$(\alpha, 0)$ partition with a PD$(\beta, 0)$ partition gives a PD$(\alpha \beta, 0)$ partition (Pitman 2002; Bolthausen and Sznitman 1998; Bertoin and Le Gall 2000). This observation along with Lemmas 5.6, 5.7, and 5.8 give Theorem 3.1.

**Lemma 5.7** For $t > T_{e-1}$ and $j, k = 1, 2, \ldots, n$, $P(\mathcal{V}(\ell_j(t)) = v_e) \to 1$ (no mutation events occur) and $P(\ell_j(t) \neq \ell_k(t)) \to 1$ if $j \neq k$ (no coalescent events occur).

**Proof** Consider the probability that the $\ell_j$ lineage experiences a mutation at time $t > T_{e-1}$. For such an event to occur, a $v_{e-1} \rightarrow v_e$ mutation must occur, and the resultant $v_e$ variant must be in the $\ell_j$ lineage. The rate of $v_{e-1} \rightarrow v_e$ mutations is given by $\mu \gamma \mathbb{E} h_{v_{e-1}}(t)$ which is trivially bounded by $O(\mu \mathbb{E})$. By symmetry the $v_e$ variant resulting from a mutation is in $\ell_j$ with probability $O\left(\frac{1}{\mathbb{E}}\right)$. By Proposition 5.1, for $t > T_{e-1}$, this probability is bounded above by $O\left(\frac{1}{\mu \mathbb{E}^2}\right)$. From this we have:

$$P(\text{no mutation event on } [T_{e-1}, T_{\text{sample}}]) = \exp\left[-\int_{T_{e-1}}^{T_{\text{sample}}} ds \mathcal{O}\left(\frac{1}{\mu \mathbb{E}^2}\right)\right] = \exp\left[-\mathcal{O}\left(\frac{1}{\mu \mathbb{E}^2}\right)\right] \to 1. \quad (45)$$

In the last line above, we have used the result $T_{\text{sample}} - T_{e-1} = O(\delta)$. To see this, note that after $T_{e-1}$, the $v_e$ variants expand deterministically. Arguments similar to those used in Proposition 5.1 show that $v_e$ will push $v_{e-1}$ to $O(\delta)$ levels in $O(\delta)$ time.
The argument for no coalescent events is similar. For $\ell_j, \ell_k$ to coalesce at time $t$, a $v_e$ variant must give birth to a new $v_e$ child cell, which occurs with rate $O(\mathbb{E}e(t))$ and $\ell_j(t), \ell_k(t)$ must be, in no particular order, precisely these parent and child cells, which occurs with probability $O(1/(\mathbb{E}e(t))^2)$. This leads to

$$P(\text{no coalescent event on } [T_{e-2}, T_{\text{sample}}]) = \exp\left[-\int_{T_{e-1}}^{T_{\text{sample}}} ds O\left(\frac{1}{\mathbb{E}e(s)}\right)\right] = \exp\left[-O\left(\frac{\delta}{\mu^2 \mathbb{E}^2 \delta^2}\right)\right] \to 1. \quad (46)$$

As mentioned, Lemma 5.7 allows us to consider $[T_i, T_{i+1})$ under the assumption that all lineages are of type $v_{i+2}$ at $T_{i+1}$. With this in mind, we introduce the following definitions:

$$\tau_j = \inf\{t : \mathcal{V}(\ell_j(t)) = v_{i+2}\}, \quad a_j = \ell_j(\tau_j). \quad (47)$$

$\tau_j$ is the time of the $v_{i+1} \rightarrow v_{i+2}$ mutation event on $\ell_j$, and $a_j$ is the specific infected cell, a $v_{i+1}$ variant, that produces the cell $\ell_j(\tau_j)$, a $v_{i+2}$ variant. We use $a_j$ as a mnemonic for “ancestor” since $a_j$ will be the ancestor of $\ell_j(T_{i+1})$.

**Lemma 5.8** Let $j, k \in \{1, 2, \ldots, n\}$ and assume that $\mathcal{V}(\ell_j(T_{i+1})) = v_{i+2}$ for all $j$. Then $P(\tau_j \in [T_i, T_{i+1}]) \to 1$ (a $v_{i+1} \rightarrow v_{i+2}$ mutation occurs on $[T_i, T_{i+1}]$) and $P(\mathcal{V}(\ell_j(T_i)) = v_{i+1}) \to 1$. Further, for $j \neq k$,

$$P(\ell_j(T_i) = \ell_k(T_i) \text{ if and only if } a_j = a_k) \to 1. \quad (48)$$

**Proof** Proposition 5.1 shows that $P(\mathcal{V}(\ell_j(T_i)) \neq v_{i+2}) \to 1$. This immediately implies that $P(\tau_j \in [T_i, T_{i+1}]) \to 1$. For $t \in [T_i, \tau_j]$, essentially the same arguments that gave Lemma 5.7 show that in the SPL $v_{i+1}$ variant lineages experience mutation events prior to $T_i$ with probability 0. Consequently, we can conclude $P(\mathcal{V}(\ell_j(T_i)) = v_{i+1}) \to 1$.

Now we consider coalescent events. If $\tau_j = \tau_k$, then we have $a_j = a_k$ as required by the lemma. So now assuming that $\tau_j > \tau_k$, we want to show that in the SPL two lineages coalesce prior to $T_i$ with probability 0. Since $\tau_j \neq \tau_k$, $\ell_j$ and $\ell_k$ cannot coalesce during $[\tau_j, T_{i+1}]$, otherwise we would necessarily have $\tau_j = \tau_k$. Further, since $\ell_j$ and $\ell_k$ are of different variant type during $(\tau_k, \tau_j)$, no coalescent event occurs on $(\tau_k, \tau_j]$. On the interval $[T_i, \tau_k)$, $\ell_k$ and $\ell_j$ are of type $v_{i+1}$, and the same arguments that gave Lemma 5.7 show that $v_{i+1}$ variants do not coalesce prior to $T_i$ with probability approaching 1. We are left with the possibility of a coalescent event at time $\tau_k$. This would mean that $\ell_j(\tau_k-)$, which is of type $v_{i+1}$, produces a mutant child cell that is of type $v_{i+2}$, which is precisely $\ell_k(\tau_k)$. However, the mutation event associated with $\ell_k(\tau_k)$ is equally likely to be produced by any variant $v_{i+1}$ at time $\tau_k$. The probability that $\ell_j(\tau_k-)$ is the cell chosen is $\frac{1}{\mathbb{E}e_{i+1}(\tau_k)}$, which is bounded as

$\mathbb{S}$ Springer
follows:
\[
\frac{1}{\mathbb{E} e_{i+1}(\tau_k)} < \frac{1}{\mathbb{E} e_{i+1}(T_i)} \to \frac{1}{O(\mu^2 E^2 \delta^2)} \to 0.
\]
(49)

\(\square\)

6 Full Escape Graph

In this section we generalize the arguments used in Sect. 5 for the linear escape graph to the full escape graph. The arguments are similar, so we emphasize the novel ideas needed for the full escape graph case. As we did for the linear escape graph, we divide the dynamics into time intervals \([T_i, T_{i+1}]\). For the linear escape graph, the \(i\)th variant class, \(E_i\), is composed of a single variant \(v_i\), and the \(T_i\) are defined by \(e_i(T_i) = \delta\) in (25). To generalize this definition to the full escape graph, we let \(T_i\) be the first time any of the variant populations in class \(i\) reaches a scaled population size \(\delta\):

\[
T_i = \inf\{t : \exists v \in E_i \text{ such that } e_v(t) \geq \delta\}.
\]
(50)

During \([T_i, T_{i+1}]\), for the linear escape graph, only \(v_{i+1}\) variants spawn \(v_{i+2}\) variants. Recalling that \(P(v)\) is the set of variant types that can mutate into variant \(v \in E_{i+2}\), for the full escape graph, all \(v' \in P(v)\) spawn \(v\) variants. For example, \(1100\) can be spawned by \(1000\) or \(0100\). Consequently, the pop value result, (30), is generalized as follows:

\[
e_v(T_{i+1}) \mu E \delta \to D_v,
\]
(51)

where the \(D_v\) and the associated \(D_{v' \rightarrow v}, S_{v' \rightarrow v}\) are given in Definition 3.2. By (51) a generalized version of Proposition 5.1 holds for the full escape graph. Namely, during \([T_i, T_{i+1}]\) variants in \(E_i\) sweep to dominance, pushing the \(E_{i-1}\) variants to \(O(\delta)\) levels. Concurrently, \(E_{i+1}\) variants rise to \(O(\delta)\) levels while spawning \(E_{i+2}\) variants. Spawning events occur for every \(v \in E_{i+2}, v' \in P(v)\) pair, and each \(v' \rightarrow v\) combination is associated with the value \(D_{v' \rightarrow v}\). Proposition 5.1 is generalized by including all such \(D_{v' \rightarrow v}\) values as described in Definition 3.2.

The factor \((D_{v'}/D_{\max, i+1})^2\) in the expression for \(D_{v' \rightarrow v}\) is not present in (30). As this difference is the main technical novelty in moving from a linear to a full escape graph, we focus on its derivation. Assume that at time \(T_i\), all variants \(v' \in E_{i+1}\) have \(e_v(T_{i})/(\mu E^2 \delta^2) \to D_v\). Heuristically, since \(\mathbb{E} e_v(T_{i}) \to \infty\), the \(e_v\) dynamics are deterministic in the SPL, and we have

\[
e_{v'}(t) \mu E \delta \to D_{v'} H_{v'}(t),
\]
(52)

where

\[
H_{v'}(t) = \exp\left[\int_{T_{i}}^{t} ds (\gamma h(s) - k_{i+1})\right].
\]
(53)
Notice that \( H_v'(t) \) is dependent only on the variant class \( E_{i+1} \) through \( k_{i+1} \) but not on the specific variant \( v' \) in \( E_{i+1} \). This leads to the ratio for \( v', v'' \in E_{i+1} \),

\[
\frac{e_{v'}(T_{i+1})}{e_{v''}(T_{i+1})} \to \frac{D_{v'}}{D_{v''}}. \tag{54}
\]

Recall that \( T_{i+1} \) is defined as the first time for which some variant \( v' \in E_{i+1} \) is of scaled population size \( \delta \). Let \( v_{\text{max},i+1} \) be that variant. Then, by definition

\[
e_{v_{\text{max},i+1}}(T_{i+1}) = \delta. \tag{55}
\]

Further, from (52) we know that \( D_{v_{\text{max},i+1}} = D_{\text{max},i+1} \). Plugging the equation directly above into (54) with \( v'' = v_{\text{max},i+1} \) gives

\[
\frac{e_{v'}(T_{i+1})}{\delta} \to \frac{D_{v'}}{D_{\text{max},i+1}}. \tag{56}
\]

Now consider the \( v \in E_{i+2} \) for which \( v' \in P(v) \). The arguments of Sect. 5 show that when \( e_{v'}(T_{i+1}) = \delta \), then \( e_{v'}(T_{i+1})/\mu^2 \delta^2 \to S \). But for the full escape graph, \( e_{v'}(T_{i+1}) \) is given by (56), and consequently we must replace \( \delta \) by \( (D_{v'/v_{\text{max},i+1}}) \delta \) in (30). This substitution gives the formula for \( D_{v' \to v} \) and, in turn, (51).

Lemmas 5.7 and 5.8 proved for the linear escape graph apply to the full escape graph with almost identical proofs. Lemma 5.6 requires modification. Recalling the notation of Sect. 5.2, consider \( n \) lineages, \( \ell_j \) for \( j = 1, 2, \ldots, n \), sampled at \( T_{i+1} \) and assume that \( V(\ell_j(T_{i+1})) = v \in E_{i+2} \). Decompose \( e_{v}(T_{i+1}) \) by

\[
e_{v}(T_{i+1}) = \sum_{v' \in P(v)} e_{v' \to v}(T_{i+1}), \tag{57}
\]

where \( e_{v' \to v} \) corresponds to \( v \) variants at \( T_{i+1} \) that descend from a \( v' \to v \) mutation. From the discussion following (51) we have

\[
P(\ell_j(T_{i+1}) \text{ descends from a } v' \to v \text{ mutation}) = \frac{e_{v' \to v}(T_{i+1})}{e_{v}(T_{i+1})} \to \frac{D_{v' \to v}}{D_{v}}. \tag{58}
\]

In the case of an escape graph, we do not need to know the pop values in order to construct the genealogies. Importantly, then, genealogies are not conditioned on a specific realization of the dynamics. However, in the full escape graph case, through the \( D_{v} \) in Definition 3.2, we build genealogies while conditioning on the pop values. In turn, we need to condition the PD(1/2, 0) on the pop values. By standard rescaling methods for Poisson–Dirichlet distributions, this is equivalent to conditioning PD(1/2, 0) on the \( S_{v' \to v} \) (Pitman 2002).

Lineages for which \( V(\ell(T_{i})) = v \) and \( V(\ell(T_{i+1})) = v' \) are contained in the \( U_{v',v} \) of Definition 3.3. Lemma 5.6, modified to consider PD(1/2, 0 | \( S_{v' \to v} \)) instead of PD(1/2, 0), can be applied to each \( U_{v',v} \) separately resulting in the partitions \( U_{v',v}^{\text{PD}} \). Concatenating this procedure through the \([T_i, T_{i+1}] \) gives Theorem 3.4.
The Coalescence of Intrahost HIV Lineages Under Symmetric CTL Attack

Acknowledgements

I thank two anonymous reviewers for comments and suggestions that greatly improved this paper. In particular, one of the reviewers conjectured that the linear escape graph genealogies could be described through the Poisson–Dirichlet distribution. This conjecture led me to the current form of Theorems 3.1 and 3.2. In an earlier version of the paper, these theorems were expressed in a different form. I express my deep gratitude to this reviewer for her/his assistance.

Appendix

A.1 Proof of Lemma 5.2

In this section, we provide the technical details that support Conclusions 3 and 4 of Lemma 5.2. From Conclusions 1 and 2, we can reduce (12) to the following:

\[
\frac{dh}{dt} = g\left(1 - h - h(e_i + e_{i-1} + O_P(\delta))\right),
\]

\[
\frac{de_i}{dt} = \gamma e_i \left(h - \frac{k_i}{\gamma}\right) + O(\mu),
\]

\[
\frac{de_{i-1}}{dt} = \gamma e_{i-1} \left(h - \frac{k_{i-1}}{\gamma}\right) + O(\mu).
\]  

(59)

The $O(\mu)$ terms in the last two equations directly above can be ignored because $T_i^h - T_i = O(|\log(\delta)|)$ and $\mu|\log(\delta)| \to 0$. Dropping these terms, we note the following relation:

\[
\frac{d(\log(e_i) - \log(e_{i-1}))}{dt} = \Delta k.
\]  

(60)

Integrating the above equation and using our assumptions on $e_{i-1}(T_i)$ and $e_i(T_i)$, we find

\[
\frac{e_i(t)}{e_{i-1}(t)} = O_P(\delta \exp[\Delta k(t - T_i)]).
\]  

(61)

If we can show that $e_i$ is bounded, then as $t$ grows, (61) implies that $e_{i-1}$ collapses. To see that $e_i$ is bounded, set

\[
z(t) = \frac{h(t)}{g} + \frac{e_i(t)}{\gamma} + \frac{e_{i-1}(t)}{\gamma}.
\]  

(62)

Then by straightforward differentiation,

\[
\frac{dz}{dt} = 1 - h - h \cdot o(\delta) - e_i \frac{k_i}{\gamma} - e_{i-1} \frac{k_{i-1}}{\gamma}
\]

\[
\leq 1 - \min\left(g(1 - o(\delta)), k_i, k_{i-1}\right)z.
\]  

(63)

Since $z(t)$ is nonnegative, we find that $z(t)$ must be bounded. In turn, $h$, $e_i$, and $e_{i-1}$ must be bounded. Returning to (61) and setting $t \geq T_i + 2/\Delta k|\log(\delta)|$, we find, since $e_i$ is bounded,

\[
e_{i-1}(t) = O_P(\delta),
\]  

(64)
and this gives Conclusion 4.

Once \( t > 2/\Delta k|\log(\delta)| \), we can further reduce (59) to

\[
\begin{align*}
\frac{dh}{dt} &= g\left(1 - h - h(e_i + O_p(\delta))\right), \\
\frac{de_i}{dt} &= \gamma e_i \left(h - \frac{k_i}{\gamma}\right).
\end{align*}
\] (65)

Consider then (65). Ignoring the \( O_p(\delta) \) term for a moment, the system is not dependent on \( \delta \). Since we have shown \( h, e_i \) to be bounded, application of the Poincaré–Bendixon theorem shows that the system converges to its nontrivial equilibrium, \( h = \frac{k_i}{\gamma} \) and \( e_i = \frac{1-h}{h} \). Now consider the \( O_p(\delta) \) term. Given some fixed distance \( \epsilon > 0 \), if we run the system from \( t = 2/\Delta k|\log(\delta)| \) to \( t = 3/\Delta k|\log(\delta)| \), we are guaranteed by choosing \( \delta \) sufficiently small to be within \( \epsilon \) of the equilibrium. In turn, taking \( \epsilon \) small, we can linearize (65) about its equilibrium.

Straightforward computation shows that both eigenvalues of the linearized system have negative real part bounded above by

\[
\rho = -\left(g\gamma + O_p(\delta)\right) \min\left(1, \frac{4k^2}{g\gamma} \left(1 - \frac{k_i}{\gamma}\right)\right).
\] (66)

Running the system from \( t = 3/\Delta k|\log(\delta)| \) to \( t = (3 + \frac{2}{|\rho|})|\log(\delta)| \) forces (65) to within \( O_p(\delta) \) of the equilibrium. This gives Conclusion 3.

A.2 Proof of Propositions 5.4 and 5.5

We first prove Proposition 5.4. Consider the Laplace transform of \( I(T)/\omega \),

\[
E\left[\exp\left[-\lambda\frac{I(T)}{\omega}\right]\right] = \exp[-\psi],
\] (67)

where

\[
\psi = \int_{-\infty}^{T} dt \mu \exp[r_1 t] \left(1 - E[\exp\left[-\lambda\frac{B(T - t)}{\omega}\right]\right),
\] (68)

and where \( B(t) \) is a continuous-time branching process with birth and death rates \( b \) and \( d \) run for time \( t \) and given \( B(0) = 1 \).

The formula for \( E[\exp[-\lambda B(t)/\omega]] \) is well known (Athreya and Ney 1972). Plugging into this formula and shifting \( t \) leads to

\[
\psi = \left(\mu \exp[r_1 T]\right) \int_{0}^{\infty} dt \exp[-r_1 t] \\
\times \left(\frac{r_2(\exp[-\lambda/\omega] - 1)}{b(\exp[-\lambda/\omega] - 1) - \exp[-r_2 t](b \exp[-\lambda/\omega] - d)}\right).
\] (69)

Making the substitution \( s = (b/r_2) \exp[r_2 t]/\omega \) gives

\[
\psi = A_0 \int_{\frac{b}{r_2 \omega}}^{\infty} ds \frac{1}{s^{1+\alpha}} \left(1 - \frac{1}{1 + \lambda s}\right) \left(1 + O\left(\frac{1}{\omega}\right)\right),
\] (70)

\( \text{Springer} \)
where $\alpha = r_1 / r_2$ and

$$A_0 = \left( \frac{b}{r_2} \right)^{\frac{1}{\alpha}} \frac{1}{b}. \quad (71)$$

The expression $1/(1 + \lambda s)$ in (70) can be recognized as the Laplace transform of an exponential. Substituting the density of an exponential and taking the $\omega \to \infty$ limit leads to

$$\psi \to A_0 \int_0^\infty ds \frac{1}{s^{1+\alpha}} \int_0^\infty dz (1 - \exp[-\lambda s z]) \exp[-z]. \quad (72)$$

Applying the Fubini theorem, we arrive at

$$\psi \to A_1 \int_0^\infty \frac{\alpha}{\Gamma(1-\alpha)} ds \frac{1}{s^{1+\alpha}} (1 - \exp[-\lambda s]), \quad (73)$$

where $\Gamma$ is the gamma function, and

$$A_1 = \left( \frac{b}{r_2} \right)^{\frac{1}{\alpha}} \frac{1}{b} \Gamma(1 + \alpha) \Gamma(1 - \alpha) \frac{1}{\alpha}. \quad (74)$$

All of the above gives

$$E[-\lambda I(T_{\text{sample}})/\omega] \to \exp[-A_1 |\lambda|^{\frac{r_1}{r_2}}]. \quad (75)$$

The Laplace transform above is that of $A_1^{r_2/r_1} S(r_1/r_2, 1, 1)$ (Nolan 2011; Bertoin 1996), which demonstrates Proposition 5.4.

Proposition 5.5 follows almost directly from the arguments that gave (73). By the uniqueness of the Levy measure, $I(T_{\text{sample}})/\omega$ converges to a nonhomogeneous Poisson process with Levy measures proportional to $s^{-1 - r_1/r_2}$. In other words, the distribution of the jumps that form $I(T)/\omega$ converge to those of a stable process of index $r_1/r_2$. Each such jump corresponds to the descendants of an immigrant. Partitioning according to immigrants is then given by $PD(r_1/r_2, 0)$ by existing results (Pitman 2002).

A.3 Proofs of Conclusion 2 of Lemma 5.3 and Lemma 5.6

We need to frame Propositions 5.4 and 5.5 in the context of spawning period dynamics under the SPL. In the period $[T_i, T_{i+1}]$, we consider “immigrants” representing $v_{i+1} \to v_{i+2}$ mutations. We have the following facts that relate to the assumptions of Proposition 5.4:

1. The number of $v_{i+1}$ variants at $t$ is $E\delta \exp[-(\Delta k + O(\delta))(T_{i+1} - t)]$.
2. $v_{i+1} \to v_{i+2}$ mutations occur at rate $\mu E\delta \exp[-(\Delta k + O(\delta))(T_{i+1} - t)]$.
3. $v_{i+2}$ variants produced descendants according to a continuous-time binary branching process with birth and death rates $k_i + O(\delta)$ and $k_{i+2} + O(\delta)$, respectively.

The $O(\delta)$ terms in the rates above will not affect the SPL limits and can be ignored. Essentially, under the SPL, the effects of $O(\delta)$ perturbations in the rates will disappear in the integral (69). If we set $T = T_{i+1}$ and $\omega = (\mu E k_i \delta)^2$, then we may apply
Proposition 5.4 to the spawning period except that “immigrants” arrive only after time $T_i$. This does not impact the $\psi$ limit in the proof of Proposition 5.4. Indeed, (70) is altered to

$$
\psi = A_0 \int_{\frac{b}{r_2r_1^\infty}}^b \frac{1}{s^{1+\alpha}} \left( 1 - \frac{1}{1 + \lambda s} \right) \left( 1 + O\left( \frac{1}{\omega} \right) \right) \, ds,
$$

(76)

where $r_2 = 2\Delta k$, $r_1 = \Delta k$, and $\alpha = 1/2$. Under the SPL, $\omega \to \infty$, so the lower bound of integration in (76) goes to 0. For the upper bound, note that

$$
\exp\left[ r_1(T_{i+1} - T_i) \right] = \frac{e_i^{i+1}(T_{i+1})}{e_i^{i+1}(T_i)} = \frac{\delta}{O(\mu^2E\delta^2)},
$$

(77)

giving

$$
\frac{\exp[r_2(T_{i+1} - T_i)]}{\omega} = \left( \frac{1}{O(\mu^2E\delta)} \right)^2 \to O\left( \frac{1}{(\mu^3E^2\delta^2)^2} \right) \to \infty.
$$

(78)

So the upper bound of integration in (76) goes to $\infty$, and Proposition 5.5 applies. Lemma 5.6 now applies by the same arguments as Proposition 5.5.

Plugging in the appropriate values for $r_1$, $r_2$, $b$ and considering the $k_i$ factor in $\omega$, we find

$$
\frac{e_i^{i+2}(T_{i+1})}{\mu^2E\delta^2} \to \left( \frac{k_i}{\Delta k} \frac{\pi^2}{2} \right) S\left( \frac{1}{2}, 1, 1 \right).
$$

(79)

This gives Conclusion 2 of Lemma 5.3.

A.4 Justification of Deterministic Approximations

To justify our deterministic approximations, we describe the steps needed to make the results of Sect. 5.1 rigorous. Associated with the time interval $[T_i, T_{i+1}]$ we define $h^{\text{det}}$ and $e^{\text{det}}_j$ for $j \leq i + 2$ as follows:

$$
\frac{dh^{\text{det}}}{dt} = g \left( 1 - h^{\text{det}} - \sum_{j'=0}^{i+2} e^{\text{det}}_{j'}h^{\text{det}} \right).
$$

(80)

For $j \leq i$,

$$
\frac{de^{\text{det}}_j}{dt} = e^{\text{det}}_j h^{\text{det}} - k_j e^{\text{det}}_j + \mu \gamma e^{\text{det}}_{j-1} h^{\text{det}},
$$

(81)

and

$$
de^{\text{det}}_{i+1} = \begin{cases} 
ed_{i+1} & \text{if } t \leq T^h_i, \\
(e^{\text{det}}_j (\gamma h^{\text{det}} - k_j) + \mu \gamma e^{\text{det}}_i h^{\text{det}}) \, dt & \text{if } t > T^h_i,
\end{cases}
$$

(82)
\[ de_{i+2}^{\text{det}} = \begin{cases} 0 & \text{if } t \leq T_i^h, \\ \frac{1}{E}(dP(yE_{j}^{\text{det}}h^{\text{det}}) - dP(k_{j}E_{j}^{\text{det}}) + dP(\mu yE_{i+1}^{\text{det}}h^{\text{det}})) & \text{if } t > T_i^h. \end{cases} \] (83)

System (80)–(82) reflects the assumptions we make in the proofs of Lemmas 5.2 and 5.3. Namely, \( e_j \) dynamics for \( j \geq i \) are taken as deterministic, \( e_{i+1} \) dynamics are taken as deterministic only after \( T_i^h \), and \( e_{i+2} \) dynamics are analyzed stochastically on all of \([T_i, T_{i+1}].\)

Let \( \Delta h = h - h^{\text{det}} \) and \( \Delta e_j = e_j - e_j^{\text{det}} \) for \( j \leq i + 2 \). We assume that \( \Delta h(T_i) = \Delta e_j(T_i) = 0 \) for all \( j \). Set

\[ v = (\Delta e_i, \Delta e_{i-1}, \ldots, \Delta e_1, \Delta h). \] (84)

The following lemma shows that system (80)–(82) behaves very similarly to the fully stochastic system (12).

**Lemma A.1**

\[ P\left( \sup_{t \in [T_i, T_{i+1}]} \|v(t)\| > \frac{1}{E^\frac{1}{8}} \right) \to 0 \] (85)

and

\[ \sup_{t \in [T_i, T_{i+1}]} \left| \frac{e_{i+1}(t)}{e_{i+1}^{\text{det}}(t)} - 1 \right| \to 0. \] (86)

The results of Sect. 5.1 can be made rigorous by using Lemma A.1 to make straightforward modifications of the proofs. For example, to demonstrate Conclusion 2 of Lemma 5.3, we need to modify the arguments made in Sect. A.2. More precisely, instead of considering (68), we consider

\[ \psi = \int_{T_i}^{T_{i+1}} dt \mu E e_{i+1}(T_{i+1} - t) \left( 1 - E \left[ \exp \left( -\frac{\lambda}{\mu^2 E \delta^2} B(T_{i+1} - t) \right) \right] \right). \] (87)

which can be decomposed into

\[ \psi = I_1 + I_2, \] (88)

where

\[ I_1 = \int_{T_i}^{T_{i+1}} dt \mu E e_{i+1}^{\text{det}}(T_{i+1} - t) \left( 1 - E \left[ \exp \left( -\frac{\lambda}{\mu^2 E \delta^2} B(T_{i+1} - t) \right) \right] \right), \]

\[ I_2 = \int_{T_i}^{T_{i+1}} dt \mu E e_{i+1}^{\text{det}}(T_{i+1} - t) \left( e_{i+1}^{\text{det}}(T_{i+1} - t) - 1 \right) \]

\[ \times \left( 1 - E \left[ \exp \left( -\frac{\lambda}{\mu^2 E \delta^2} B(T_{i+1} - t) \right) \right] \right). \] (89)
The analysis of $I_1$ proceeds exactly as in Appendix A.2 with the modifications of Appendix A.3. $I_2 \to 0$ given (86) and the convergence result for $I_1$.

We have left the task of proving Lemma A.1. Our approach is a specific implementation of the general approach outlined in Darling and Norris (2008). See Leviyang (2011) for another example of these types of arguments in the context of viral dynamics.

**Proof** Define the process $M_j(t)$ by

\[ dM_j = de_j - (e_j(\gamma h - k_j) + \mu \gamma h e_j) dt. \tag{90} \]

$M_j(t)$ is a martingale on the filtration defined by $h, e_j, e_j^{-1}$.

We first focus on (85). Using stopping times, we force $\Delta h, \Delta e_j$ for $j \leq i$ to be small. We then linearize the equations governing $\Delta h, \Delta e_j$ and show that the stopping times may be removed while still insuring that $\Delta h, \Delta e_j$ are small.

For the subinterval $[T_i, T_i^h]$, define the following stopping times:

\[ T_A = \inf \{ t > T_i : |\Delta e_j| \geq \frac{1}{(E_1^4)} \text{ for } j \leq i \}, \]

\[ T_B = \inf \{ t > T_i : |\Delta h| \geq \frac{1}{(E_1^4)} \}, \]

\[ T_C = \inf \{ t > T_i : e_j \geq \frac{\mu^2 E(\mu E)^{\kappa}, e_j^{det} \geq \mu^2 E(\mu E)^{\kappa}}{j \geq i + 1} \}. \]

In the definition of $T_C$, $\kappa$ is any positive constant less than $1/4$. The factor $(\mu E)^{\kappa}$ ensures that $P(T_C < T_i^h) \to 0$ while maintaining sufficient control of $e_{i+1}$. Set $T = T_A \wedge T_B \wedge T_C$. The equations for $\Delta h$ and $\Delta e_j$ on $[T_i, T \wedge t]$ with $j \leq i$ are

\[ d\Delta h = -g \Delta h - g \sum_{k=1}^{n} h^{det} \Delta e_k - g \sum_{k'=1}^{n} e_{k'}^{det} \Delta h + O\left( \frac{1}{(E_1^4)^2} + \frac{\mu^2 E(\mu E)^{\kappa}}{(E_1^4)^{1/4}} \right), \tag{92} \]

\[ d\Delta e_j = \Delta e_j(\gamma h^{det} - k_j) + e_j^{det} \Delta h + dM_j + O\left( \frac{1}{(E_1^4)^2} + \frac{\mu}{(E_1^4)^{1/4}} \right). \]

Set

\[ M = (M_i, M_{i-1}, \ldots, M_1, 0). \tag{93} \]

Then (92) can be integrated:

\[ v(t \wedge T) = \int_{T_i}^{t \wedge T} dM(s) \exp \left[ \int_{s}^{t \wedge T} ds' \Sigma(s') \right] + \int_{T_i}^{t \wedge T} ds \exp \left[ \int_{s}^{t \wedge T} ds' \Sigma(s') \right] \times O\left( \frac{1}{(E_1^4)^2} + \frac{\mu^2 E(\mu E)^{\kappa}}{(E_1^4)^{1/4}} \right). \tag{94} \]
where $\Sigma$ is a matrix defined through (92) that depends on $h_{\text{det}}$ and $e_{j}^{\text{det}}$. Note that the first integral to the right of the equality directly above is a martingale.

Now, consider $t \wedge T \wedge T_i^h$. Since, by definition, $T_i^h - T_i = O(\log(\delta))$, we have

$$
\| \exp \left[ \int_s^{t \wedge T \wedge T_i^h} ds' \Sigma (s') \right] \|_\infty = O\left( \frac{1}{\delta^c} \right) \tag{95}
$$

for some constant $c$ that depends only on the parameters of (12). Then using the Doob inequality and noting that the quadratic variation of $M(t)$ is $O\left( \frac{1}{E} \right)$, we can arrive at

$$
E \left[ \sup_{s \in [T_i, T_i^h]} |v(s)|^2 \right] \leq O \left( \frac{1}{E \delta^{c_1}} \right) + O \left( \frac{1}{(\frac{1}{E})^2 \delta^{c_2}} \right) + \frac{\mu^2}{(\frac{1}{E})^4 \delta^{c_3}} + \frac{\mu^4 \delta^{2} (\mu E)^2 \kappa}{(\frac{1}{E})^2 \delta^{c_4}} \tag{96}
$$

where the $c_i$ are constants independent of the SPL. A Chebyshev bound and the limit $(\mu^2 E)^2 (\mu E)^\kappa / \delta^c \to 0$ then give

$$
P \left( \sup_{s \in [T_i, T_i^h]} |v(s)| > \frac{1}{E^{\frac{1}{4}}} \right) \to 0. \tag{97}
$$

The above limit shows that the stopping times $T_A, T_B$ may be removed from $T$ outside of a set with collapsing probability under the SPL. The arguments given in the proof of Lemma 5.2 can be used to show $P(T_C < T_i^h) \to 0$, and so we arrive at

$$
P \left( \sup_{s \in [T_i, T_i^h]} |v(s)| > \frac{1}{E^{\frac{1}{4}}} \right) \to 0. \tag{98}
$$

Now we consider the time interval $[T_i^h, T_i+1]$. Define

$$
T_A' = \inf \left\{ t > T_i^h : |\Delta e_j| \geq \frac{1}{E^{\frac{1}{4}}} \text{ for } j \leq i \right\},
$$

$$
T_B' = \inf \left\{ t > T_i^h : |\Delta h| \geq \frac{1}{E^{\frac{1}{4}}} \right\}, \tag{99}
$$

$$
T_C' = \inf \left\{ t > T_i^h : e_j \geq 2\delta, e_{j}^{\text{det}} \geq 2\delta \text{ for } j \geq i + 1 \right\}.
$$

Set $T' = T_A' \wedge T_B' \wedge T_C'$. Note that $T_C'$ is defined with respect to $2\delta$.

On $[T_i^h, T']$ the arguments of Sect. 5.1 that involve deterministic dynamics give $e_{j}^{\text{det}} = O(\delta)$ for $j < i$ and $\gamma h^{\text{det}} - k_i = O(\delta)$. For $j < i$, we have the bound

$$
d \Delta e_j = (\gamma h^{\text{det}} - k_j) \Delta e_j + O(\delta) \Delta h + dM_j + O\left( \frac{1}{(E^{\frac{1}{4}})^2} + \frac{\mu}{E^{\frac{1}{4}}} \right). \tag{100}
$$
Applying the same arguments as above and noting that $T' - T_i^h = O(1/\delta)$ gives for $j < i$,
\[
P\left( \sup_{s \in [T_i^h, T']} |\Delta e_j(s)| > \frac{1}{\mathbb{E}^{\frac{1}{\delta}}} \right) \to 0. \tag{101}
\]
We need to control $\Delta e_i, \Delta h$. It is easy to check that the $2 \times 2$ matrix that gives the linearized dynamics of $\Delta e_i, \Delta h$ on $[T_i^h, T']$ is negative definite. Then similar arguments as used above show
\[
P\left( \sup_{s \in [T_i^h, T']} |\Delta e_i(s)| > \frac{1}{\mathbb{E}^{\frac{1}{\delta}}} \right) \to 0. \tag{102}
\]
and similarly for $\Delta h$. Equation (102) allows us to remove $T'_A, T'_B$ from the definition of $T'$. As a result, we have demonstrated (85), except that $t$ is restricted to $[T_i, T'_C]$ instead of $[T_i, T_{i+1}]$.

We now turn to $e_{i+1}$ and the proof of (86). Clearly, $\Delta e_{i+1}(t) = 0$ for $t \in [T_i, T_i^h]$, and we need only consider the interval $[T_i^h, T_{i+1}]$. From the results above and the analysis of Sect. 5.1, for $t \in [T_i^h, T'_C]$, we have
\[
d\left( \frac{e_{i+1}}{e_{i+1}^{\text{det}}} \right) = \frac{e_{i+1}^{\text{det}}}{e_{i+1}^{\text{det}}} \left( O\left( \frac{1}{(\mathbb{E}^{\frac{1}{\delta}})^2} + \frac{\mu}{\mathbb{E}^{\frac{1}{\delta}}} \right) \right) + \frac{dM_{i+1}}{e_{i+1}^{\text{det}}} . \tag{104}
\]
Integrating (104) gives for $t \in [T_i^h, T'_C]$,
\[
\frac{e_{i+1}^{\text{det}}}{e_{i+1}^{\text{det}}}(t) - 1 = \left( \exp\left[ O\left( \frac{t - T_i^h}{(\mathbb{E}^{\frac{1}{\delta}})^2} + \frac{(t - T_i^h)\mu}{\mathbb{E}^{\frac{1}{\delta}}} \right) \right] - 1 \right) + \int_{T_i^h}^t dM_{i+1}(s) O\left( \frac{1}{e_{i+1}^{\text{det}}(s)} \right). \tag{105}
\]

Taking the second moment, noting that $T'_C - T_i^h = O(1/\delta)$, and using the Doob inequality give
\[
E\left[ \sup_{t \in [T_i^h, T'_C]} \left( \frac{e_{i+1}^{\text{det}}}{e_{i+1}^{\text{det}}}(t) - 1 \right)^2 \right] \leq O\left( \frac{1}{(\mathbb{E}^{\frac{1}{\delta}})^2\delta^2} + \frac{\mu^2}{(\mathbb{E}^{\frac{1}{\delta}})^2\delta^2} \right) + O\left( \int_{T_i^h}^{T'_C} ds \frac{1}{\mathbb{E}^{\frac{1}{\delta}}} \left( \frac{1}{\mu^2\mathbb{E}^{\frac{1}{\delta}}} \right) \right) \tag{106}
\]
for some constant $c$. Since $\mu^2\mathbb{E}^{2\delta c} \to \infty$ by the Chebyshev inequality, we have proved (86) except that $t \in [T_i, T'_C]$.
The lemma is proved except that we have considered \([T_i, T'_C]\) rather than \([T_i, T_{i+1}]\). But we have
\[
\frac{e_{i+1}(T'_C)}{e_{i+1}^{\text{det}}(T'_C)} \rightarrow 1, \tag{107}
\]
and so \(e_{i+1}(T_C)/2\delta \rightarrow 1\). But this implies \(P(T_{i+1} < T'_C) \rightarrow 1\), and we are done. \(\square\)

References

Asmussen, S., & Glynn, P. W. (2007). *Stochastic simulation, algorithms and analysis*. Stochastic modeling and applied probability. Berlin: Springer.

Athreya, K. B., & Ney, P. E. (1972). *Branching processes*. Berlin: Springer.

Barton, N. H., et al. (2004). Coalescence in a random environment. *Ann. Appl. Probab.*, 14, 754–785.

Bertoin, J. (1996). *Levy processes*. Cambridge tracts in mathematics.

Borrow, P. H., et al. (1994). Virus-specific cd8+ cytotoxic t-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection. *J. Virol.*, 68, 6103–6110.

Carrington, M., & O’Brien, S. J. (2003). The influence of HLA genotype on AIDS. *AIDS Annu. Rev. Med.*, 54, 535–551.

Crandall, K. A. (1999). *The evolution of HIV*. Baltimore: Johns Hopkins University Press.

Darling, R. W. R., & Norris, J. R. (2008). Differential equation approximations for Markov chains. *Probab. Surv.*, 5, 37–79.

DeFranco, A. L., Locksley, R. M., & Robertson, M. (2007). *Immunity: the immune response in infectious and inflammatory disease*. London: New Science Press.

Drummond, A. J., & Rambaut, A. (2007). Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.*, 7, 214.

Bertoin, J., & Le Gall, J.-F. (2000). The Bolthausen–Sznitman coalescent and the genealogy of continuous-state branching processes. *Probab. Theory Relat. Fields*, 117, 249–266.

Bolthausen, E., & Sznitman, A.-S. (1998). On Ruelle’s probability cascades and an abstract cavity method. *Commun. Math. Phys.*, 197, 247–286.

Leigh-Brown, A. J. (1997). Analysis of HIV-1 env gene sequences reveals evidence for a low effective number in the viral population. *Proc. Natl. Acad. Sci. USA*, 94, 1862–1865.

Chun, T.-W., et al. (1997). Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature*, 387, 183–188.

Desai, M. M., & Fisher, D. S. (2007). Beneficial mutation-selection balance and the effect of linkage on positive selection. *Genetics*, 176, 1759–1798.

Drummond, A. J., & Rodrigo, A. G. (2000). Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA. *Mol. Biol. Evol.*, 17, 1807–1815.

Durrett, R., et al. (2009). A waiting time problem arising from the study of multi-stage carcinogenesis. *Ann. Appl. Probab.*, 19(2), 676–718.

Durrett, R., & Schweinsberg, J. (2004). Approximating selective sweeps. *Theor. Popul. Biol.*, 66, 129–138.

Goonetilleke, N., 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(6), 1253–1272.

Hermisson, J., & Pennings, P. S. (2005). Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*, 169, 2335–2352.

Iwasa, Y., et al. (2005). Population genetics of tumor suppressor genes. *J. Theor. Biol.*, 233, 15–23.

Kaplan, N. L., et al. (1988). The coalescent process in models with selection. *Genetics*, 120, 819–829.

Kelleher, A. D., et al. (2001). Clustered mutations in HIV-1 gag are consistently required for escape from hla-b27-restricted cytotoxic t lymphocyte responses. *J. Exp. Med.*, 193, 375–386.

Kepler, T. B., & Oprea, M. (2001). Improved inference of mutation rates: I. An integral representation for the Luria–Delbruck distribution. *Theor. Popul. Biol.*, 59, 41–48.

Koup, R. A., et al. (1994). Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome. *J. J. Virol.*, 68, 4650–4655.

Kouyos, R. D., et al. (2006). Stochastic or deterministic: what is the effective population size of HIV-1. *Trends Microbiol.*, 14(12), 507–511.
