Published in final edited form as:
J Theor Biol. 2013 July 21; 329: 82–93. doi:10.1016/j.jtbi.2013.03.026.

Modeling sequence evolution in HIV-1 infection with recombination

Elena E. Giorgi[a,b], **Bette T. Korber**[a,c], **Alan S. Perelson**[a,c], and **Tanmoy Bhattacharya**[a,c]

[a]Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM, 87545

[b]Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM, 87545

[c]Santa Fe Institute, Santa Fe, NM, 87501

Abstract

Previously we proposed two simplified models of early HIV-1 evolution. Both showed that under a model of neutral evolution and exponential growth, the mean Hamming distance (HD) between genetic sequences grows linearly with time. In this paper we describe a more realistic continuous-time, age-dependent mathematical model of infection and viral replication, and show through simulations that even in this more complex description, the mean Hamming distance grows linearly with time. This remains unchanged when we introduce recombination, though the confidence intervals of the mean HD obtained ignoring recombination are overly conservative.

Keywords

HIV; population dynamics; viral evolution

1. Introduction

The theory of population genetics has been applied to a vast number of organisms, from primates to bacteria and virus. Combined with coalescent theory and phylogenetic methods, it can trace back in time the dynamics and evolution of species. Previously (Lee et al., 2009; Giorgi et al., 2010), we developed an intra-host evolutionary model for HIV-1 during early infection, and tested the genetic diversity of early HIV-1 samples against a null model of neutral evolution. The method allows one to distinguish infections that are established by a single virus from those initiated by multiple viruses. It also allows one to estimate the time since infection and the onset of immune selection. Though the model is now widely used and performs well on different datasets (Keele et al., 2008; Wood et al., 2009; Keele et al., 2009), including very large ones obtained through 454 sequencing (Fischer et al., 2010), the viral lifecycle described in Lee et al. (2009) was a simplification of the underlying biology. Here we study a more realistic model of HIV replication and how the new approach affects the results.

Following in Sewall Wright's footsteps, Kimura (1968) used the Fokker-Planck equation to develop a continuous time model of population genetics. This was later expanded by Gillespie (1984) who developed explicit simulation frameworks for continuous time stochastic models in population genetics.

In the first part of this paper, we develop a continuous time, age-dependent model to describe the genetic drift in a growing viral population. We use a deterministic approach and derive a model using the Von Foerster equation and renewal condition with age-dependent birth and death rates. We then expand this into a completely stochastic model using the Fokker-Planck equation and follow, through simulation, the variation in genetic distances across a finite population of viral sequences. We measure genetic distances in the sample using Hamming distances (HD), defined as the number of base positions at which two sequences differ. Model simulations show that even in this fully randomized model, the growth in mean HD, averaged over multiple runs, is linear in time, and that different choices of birth and death rates only affect the rate of the growth but not this behavior.

In the second part of the paper we study how recombination affects the results. Genetic recombination takes place when a newly constructed nucleic acid molecule arises from multiple template strains. Recombination events in HIV-1 are common (Onafuwa-Nuga et al., 2009; Robertson et al., 1995; Wooley et al., 1997; Sabino et al., 1994) and result from multiple infections of the same target cell (Chen et al., 2005; Dang et al., 2004; Jung et al., 2002; Levy et al., 2009). A number of studies have indicated that the HIV recombination rate is several times larger than its mutation rate (Batorsky et al., 2011; Jetzt et al., 2000; Levy et al., 2009; Shriner et al., 2004; Suryavanshi et al., 2007; Neher et al., 2010; Zhang et al., 2010), and that recombinants reach fixation and undergo a rapid expansion in the population, thus representing an evolutionary shortcut (Ramirez et al., 2008).

Previous recombination studies have suggested that recombination affects a virus's ability to escape CTL responses (Mostowy et al., 2011) and has a significant impact on the emergence of drug resistance (Althaus et al., 2005; Fraser, 2005; Kouyos et al., 2009; Moutouhet al., 1996; Vijay et al., 2008). In Vijay et al. (2008) the authors use a simulation to study the effects of recombination on the changes in mean Hamming distance with time. However, in the works cited above, it is assumed that the host's immune response plays a major role in defining the fitness of the viral strains. In this paper, we address recombination in our original model of exponential viral growth with no positive selection pressure from the environment, a scenario that is relevant to early dynamics of viral growth in the acute phase of HIV infection prior to the initiation of an effective adaptive immune response. Under this scenario, we see that recombination does not affect the linearity in time of the mean HD growth, though it does change the growth of the variance of the HD distribution.

2. Results

2.1. Continuous time model of viral evolution

In Lee et al. (2009) we described two simplified models of viral evolution, where cell infection and cell death happen at fixed, discrete times. In this section we present a more realistic model based on a continuous time birth and death process. We also introduce age-dependency in the model, where *age* represents the time since a cell was infected. The corresponding stochastic formulation is presented in Appendix A.

As stated in Lee et al. (2009), we assume that one unique genetic strain, called the transmitted/founder virus, initiates the infection. In general, it takes some time before the host's immune response kicks in, and since during the early phase of the infection the viral population is much smaller than the target cell population, we assume that viral evolution is initially driven by exponential growth and random accumulation of mutations at allowed sites (Ribeiro et al., 2010). Genetic diversity can also be achieved through recombination, which takes place when two or more distinct strains infect one cell and give rise to new, recombined genomes. However, when the viral population is sufficiently small with respect to the target cell population, the effect of recombination during this initial phase of the

infection should be negligible. Under these hypotheses, mutations from the founder strain accumulate randomly following a Poisson distribution (Lee et al., 2009).

In all our analyses, we ignore the presence of proteins such as APOBEC, which results in hypermutated sequences and can modify the mutation rate. We have included the capability to identify and exclude APOBEC mediated hypermutation in previous implementations of our model (Giorgi et al., 2010).

Once a virion enters a cell, the HIV enzyme reverse transcriptase transcribes the viral RNA into DNA, which is then integrated into the host genome. A viral genome integrated into the DNA of the host cell is called a provirus. The time from viral entry into a target cell until the first virions start budding out (called the eclipse phase) has been estimated to last about 24 hours (Perelson et al., 1996; Markowitz et al., 2003), although in some cases the virus can lie dormant inside the cell for years. The infected cell will on average survive another 24 hours while producing virus (Markowitz et al., 2003), and during this time it will produce tens of thousands of new viral particles (Chen et al., 2007). HIV is rapidly cleared from circulation, with an average half-life of roughly 45 minutes (De Boer et al., 2010; Ramratnam et al., 1999), and on average, in early infection, between 6 and 10 virions go on to successfully infect new cells (Ribeiro et al., 2010). The number of virions that successfully infect new cells in this initial phase is called the basic reproductive ratio, usually denoted R_0.

In all that follows, we neglect the time a virus spends outside a cell and choose to follow provirus instead. We only consider mutations that may occur during reverse transcription, and neglect the extremely rare mutations that happen after integration in the cell's genome.

Given this framework, we let $I(a, g, t)$ be the number of infected cells that at time t are of age a (i.e. a represents the time elapsed since the cell was infected) and generation g (i.e. the genome it carries has undergone g infection cycles since the transmitted virus, with one reverse transcription event occurring at each infection). Here a and t are non-negative, real variables, and g is a non-negative integer.

We assume that both the birth rate $\alpha(a)$ (the number of cells infected by each cell of age a per unit time) and death rate $\beta(a)$ (the number of cells of age a that die per unit time) are functions that depend only on the age of the infected cell, a, and not on time t. Notice that $\alpha(a)$ is in fact an infection rate. Here we choose to call it birth rate to show that the classic birth and death process provides a good infection model when neglecting the time the virus spends outside the cell. Let Λ denote the lag before new virions start budding out of the infected cell (i.e. the length of the eclipse phase) and impose that $\alpha(a) = 0$ for $a \leq \Lambda$. In the discussion that follows, we implicitly assume that all functions evaluate to zero on negative arguments.

It has been shown (Von Foerster, 1959; Nisbet et al., 1982; Huddleston et al., 1983) that the dynamics of an age-structured birth and death system is determined by the following differential equations, called the Von Foerster equation and the renewal condition, respectively:

$$\begin{aligned} \frac{\partial}{\partial t} I(a, g, t) + \frac{\partial}{\partial a} I(a, g, t) &= -\beta(a) I(a, g, t) \\ I(0, g+1, t) &= \int_0^t I(a, g, t) \alpha(a) da \end{aligned} \quad (1)$$

with initial conditions

$$I(a,g,0)=\begin{cases} I_0 & \text{if } a=g=0 \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

The first equation in (1) describes how infected cells are lost via death, whereas the second one describes how newly infected cells are formed via births. This is a classic age-structured birth and death system (Murray, 2002; Kevorkian, 2002), with the addition of the parameter g for generation cycles. Such models are called age-maturity structured and have been described in the context of cell kinetics, where the parameter g in that case represents the number of cell divisions (Bernard et al., 2003; Zilman et al., 2010).

Let $f(a)=\alpha(a)\,\mathrm{e}^{-\int_0^a \beta(s)ds}$. Notice that $\mathrm{e}^{-\int_0^a \beta(s)ds}$ can be interpreted as the expected fraction of cells whose age at death is greater than a. Therefore, $f(a)$ represents the expected rate at which an infected cell survives up to age a and starts new infections. Let $\tilde{f}$ be the Fourier transform of f, i.e.

$$\tilde{f}(k)=\int_\Lambda^\infty f(a)\,\mathrm{e}^{ika}\,da. \quad (3)$$

We assume that $\beta(a)$ is a strictly positive function with $\lim_{a\to\infty}\beta(a)\neq 0$ and such that $\lim_{a\to\infty} f(a) = 0$. It follows (see Appendix B) that a closed solution of Eq. (1) is given by

$$I(g,t)=\frac{I_0}{2\pi}\int_0^t e^{-\int_0^a \beta(s)ds}\int_{-\infty}^{\infty}\tilde{f}(k)^g\,\mathrm{e}^{-ik(t-a)}\,dk\,da \quad (4)$$

2.2. Mean Hamming Distance

In Keele et al. (2008) we attained plasma samples from HIV-1 subjects early after infection, from which viral sequences were derived. Roughly 80% of those subjects were infected by a single viral strain, and we called those infections homogeneous. Under our model of neutral exponential growth, mutations from the founder strain accumulate randomly following a Poisson distribution due to reverse transcription errors introduced when the virus infects a cell. As mutations accumulate, the mean of the Poisson distribution grows with time. Therefore, given a sample of viral sequences taken from a homogeneously infected subject, it is possible to count the number of mutations across all pairs of sequences, fit a Poisson distribution to the frequency counts, and use the mean of the Poisson to estimate how much time has elapsed since the beginning of the infection.

The Hamming distance (HD) between two genetic sequences is defined as the number of bases at which the two differ. In our original model (Lee et al., 2009; Giorgi et al., 2010), we showed that for homogeneous infections, in the absence of selection, the mean HD is given by

$$E[HD](t)=\sum_d dP(HD=d|t)\approx\sum_d\frac{\sum_g I(g,t)P_{Binom}(d;2gN_B,\varepsilon)}{\sum_g I_g(t)} \quad (5)$$

where ε is the viral mutation rate per base per generation, N_B is the length of the sequences, and g is the number of generation steps since the most recent common ancestor, or MRCA. $P_{Binom}(d; n, p)$ is the binomial distribution with parameters n and p:

$$P_{Binom}(d;n,p)=\binom{n}{d}p^d(1-p)^{n-d}.$$

Here we are neglecting the possibility of multiple mutations at the same site, as these are expected to be very rare in early HIV infection (Lee et al., 2009).

In the synchronous infection model described in Lee et al. (2009), we assume that each provirus produces R_0 new proviruses and that all infected cells are of the same generation and are born and die synchronously. We calculated Eq. (5) for $I(g,t)=R_0^g$ (synchronous model) and for a slightly more complicated, but still simplified, scenario in which cells produce virions in two synchronous bursts, as a first approximation to an asynchronous infection. In both cases we saw that the mean *HD* (calculated over all pairs of sequences in a given sample) grows linearly with time and is proportional to $2g\varepsilon N_B$. In Lee et al. (2009) we showed that the MRCA is at most a couple of generations away from the transmitted/founder strain, and it coincides with the transmitted/founder strain in most cases. Since, in HIV infections, one replication cycle is roughly two days (Markowitz et al., 2003), we can use this result to estimate the time since the MRCA in early homogeneous infections (Keele et al., 2008; Wood et al., 2009; Keele et al., 2009).

We now show that the linearity in time of the mean *HD* growth remains unchanged when substituting the more general expression of $I(g, t)$ from Eq. (4) into Eq. (5). We first derive the expression of the mean HD at generation g when the death rate is constant, i.e. $\beta(a) = \beta_0$ for all $a > 0$, and the birth rate is a piecewise linear function in age of the form $\alpha(a) = 0$ for all $a \leq \Lambda$, and $\alpha(a) = \alpha_0(a - \Lambda)$ for all $a > \Lambda$. Expanding Eq. (4) using these particular birth and death rates, one obtains

$$I(g,t)=I_0\frac{\alpha_0^g}{(2g)!}\left[(t-\Lambda g)^{2g}-(\Lambda g)^{2g}\right]e^{-\beta_0 t} \quad (6)$$

Given the biological properties of HIV infection, we can impose the condition that both β and α be positive, analytic, and asymptotically constant functions, and, in addition, that $\alpha(a)$ be null both at 0 and at infinity. We can then extend the above derivation of $I(g, t)$ for general death and birth rates by noting that a bounded death rate function $\beta(a)$ can always be approximated with a step function (a function that is constant on a given partition), and $\tilde{I}$ in Eq. (4) can be calculated for a generic birth rate from Eq. (3) using the Taylor expansion of α.

Finally, we use a simulation based on equations (4) and (5) to generate $I(g, t)$ for different choices of piecewise linear death rate functions and a piecewise linear birth rate function as defined above. For these choices of birth and death rate functions, we generate the full population of infected cells, out of which we randomly sample N_0 sequences at each day since the start of infection. At any given time t, we assume that each infected cell in $I(g, t)$ carries one proviral strain that has undergone exactly g infection cycles from the MRCA. Therefore, for each one of the N_0 sequences, knowing how many infection cycles it has undergone, we randomly draw from a Poisson distribution the number of mutations it has accumulated from the MRCA. We then calculate the mean intersequence *HD* assuming a star phylogeny, in other words, for any two given sequences s_1 and s_2 in the sample, $HD[s_1, s_2] = HD_0[s_0, s_1] + HD_0[s_0, s_1]$, where s_0 is the MRCA, and HD_0 is the Hamming distance from the MRCA. Finally, we calculate the mean *HD* over all pairs and average over 10,000 runs. In Figure 1 we show the results when both death and birth rates are linear functions of

age. Different choices of birth and death rates do not change the linearity in growth of the mean *HD*, but only affect the rate at which the mean *HD* grows (results not shown). This result confirms what we have previously shown in Lee et al. (2009) using a discrete generation model, namely that the mean *HD* is proportional to $2\varepsilon g N_B$.

2.3. Recombination

We now discuss the effect of recombination on the growth of the mean HD. We start with the same scenario described in Lee et al. (2009): a single infecting strain establishes infection, no selection, exponentially growing viral population, and a negligible probability of observing doubly mutated sites (we show in Lee et al. (2009) that the probability of back mutations is $\leq O\left(\frac{g^2\varepsilon^2}{N_B}\right)$, which is negligible in our scenario of early infections). This is an apt description of early, heterosexually transmitted, HIV infections (Keele et al., 2008), and, under this scenario, prior to the onset of selection, the HD frequency counts follow a Poisson distribution. However, while in a Poisson model the mean HD of a sample of N sequences should approximately equal the HD variance times $\frac{N}{N-1}$, the HD variance was 4.7% lower than expected when averaging across all homogeneous subjects described in Keele et al. (2008). This difference fell within the 95% CI and was not statistically significant. Nonetheless, given the current advances in sequencing technology and the availability of larger sample sizes (i.e. 454 sequencing (Margulies, 2005)) it's interesting to investigate whether this difference is due to recombination, which we neglected in our original model.

We now show that in the presence of recombination the HD frequency counts no longer follow a Poisson distribution, thus explaining possible divergences between the variance and the mean. As discussed in the previous section, in the absence of recombination, our simulation results show that both the continuous and stochastic models yield the same basic results as the previous simplified discrete generation model, namely that the mean *HD* is proportional to $2\varepsilon g N_B$.

Preliminary phylogenetic analyses conducted on viral samples attained at serial time points from early, homogeneous infections suggest that the mutation rate ε may vary greatly across patients (Wallstrom et al., 2010). Therefore, in all that follows, we revert to our simplified, synchronous infection model and assume that changes in the mean *HD* growth rate caused by possible divergences from such a simplified model are folded into the parameter ε, which needs to be measured in each patient individually.

To allow for the possibility of multiple strains infecting the same cell and possibly giving rise to a recombinant strain, we introduce a recombination rate $0 < \rho < 1$. As a simplification, at each generation we let each viral strain either undergo a mutation event or a recombination (not both) with probabilities ρ and $(1 - \rho)\varepsilon$ respectively. The more general case can easily be solved using the same methods. We assume ρ is constant in time and, for simplicity, we also assume that the parental strains recombine at a single position θ, which is a uniformly distributed random variable between 0 and the length N_B of the genome. Given the MRCA s_0, let $HD_0(s, t)$ be the Hamming distance of sequence s from s_0, where s was sampled at time t. For sufficiently small sample sizes, early homogeneous infections follow a star-like phylogeny and both the HD_0 and the intersequence *HD* follow a Poisson distribution (Lee et al., 2009; Keele et al., 2008). In particular, the expectations are related, i.e. $E[HD] = 2E[HD_0]$. Under these assumptions, the HD_0 probability distribution at time $t + dt$ is given by

$$\tilde{P}(HD_0=d;t+dt)=(1-\rho)P_\varepsilon(HD_0=d;t+dt)+\rho P_\rho(HD_0=d;t+dt) \quad (7)$$

where P_ε and P_ρ are the probability of a genome being at distance d from s_0 after a mutation or a recombination event, respectively. Suppose that P_ρ has mean μ_ρ and variance σ_ρ^2. Similarly, we denote by μ_ε and σ_ε^2 the mean and variance of P_ε. Then

$$\tilde{\mu}(t)=E(HD_0|t)=(1-\rho)\mu_\varepsilon(t)+\rho\mu_\rho(t) \quad (8)$$

and

$$\tilde{\sigma}^2=Var(HD_0)=\sigma_\varepsilon^2+2\rho\mu_\varepsilon(\mu_\varepsilon-\mu_\rho)+\rho(\mu_{2,\rho}-\mu_{2,\varepsilon})-\rho^2(\mu_\varepsilon-\mu_\rho)^2 \quad (9)$$

where $\mu_{2,\varepsilon}$ and $\mu_{2,\rho}$ indicate the second moments of P_ε and P_ρ, respectively. When $\rho = 0$, we get $\tilde{\sigma}^2=\sigma_\varepsilon^2$ and when $\rho = 1$, $\tilde{\sigma}^2=\sigma_\rho^2$. Notice also that when $\mu_\varepsilon = \mu_\rho$ one has

$$\tilde{\sigma}^2(t)=\sigma_\varepsilon^2(t)+\rho(\mu_{2,\rho}(t)-\mu_{2,\varepsilon}(t)) \quad (10)$$

In what follows, we suppress the variable t when it is clear from the context. When viral genomes undergo mutation

$$P_\varepsilon(HD_0=d;t+dt)=\sum_{k=0}^{d}\tilde{P}(d-k;t)P_m(k;dt) \quad (11)$$

where $P_m(k; dt)$ is the probability of having k mutations appear in a time interval dt. Assuming the independence of $\tilde{P}$ and P_m, we can use the fact that cumulants add under convolution. Let $d\lambda$ be the mean of the distribution P_m in the time interval dt (for example, in a discrete, synchronous model, it would be the mutation rate times the sequence length). Then, omitting $O(d\lambda^2)$ terms,

$$\begin{aligned}\mu_\varepsilon(t+dt) &= \tilde{\mu}(t)+d\lambda \\ \sigma_\varepsilon^2(t+dt) &= \tilde{\sigma}^2(t)+d\lambda \\ \mu_{2,\varepsilon}(t+dt) &= \tilde{\mu}_2(t)+(1+2\tilde{\mu}(t))d\lambda\end{aligned} \quad (12)$$

2.4. Recombination alone

In order to compute P_ρ, it is useful to first envision a scenario in which an initial, genetically diverse population of I_0 sequences is allowed to grow exponentially with no accumulation of mutations. In other words, we set $\varepsilon = 0$ and $\rho = 1$. Suppose the HD_0 distribution of the I_0 sequences has initial mean μ_0 and variance σ_0^2. At each generation a recombination position θ is drawn randomly from a uniform distribution, with $\theta = 0, \ldots, N_B$, where the two parental strings split and recombine. If $\theta = 0$, then the recombinant is exactly the first sequence, whereas if $\theta = N_B$, the recombinant is the second sequence. Under this scenario it is easy to see that the mean Hamming distance of a population of sequences will remain unchanged from the initial μ_0, whereas the total variance in the population will decline. As we show in Appendix C, at any generation g, the expectation values of the mean HD_0, $\mu(g)$, and the estimator of the variance $s^2(g)$ are given by

$$\mu(g)=\mu_0 \quad (13)$$

and

$$s^2(g)=\frac{N_B-1}{N_B}\mu_0+\frac{1}{N_B}\sigma_0^2+2\frac{\sigma_0^2-\mu_0}{g+2}\left[\left(\frac{N_B}{N_B+1}\right)^{g-1}+O(N_B^{-1})\right]\prod_{j=0}^{g-1}\frac{N_j-1}{N_j} \quad (14)$$

where μ_0 and σ_0^2 are the mean and variance HD_0 of the initial population of I_0 sequences, N_j is the size of the population at generation j, and N_B is the total sequence length.

In order to test the formulae in (14), we devised a simulation based on the previously described synchronous infection model (Lee et al., 2009) where no site is mutated more than once (infinite site assumption). We start at generation $g = 0$ with I_0 sequences, each with an initial HD_0 drawn from a probability distribution with mean and variance μ_0 and σ_0^2 respectively (in the simulations shown in the figures we used $\mu_0=\sigma_0^2=10$). At each later generation we resample $I_0R_0^g$ sequences from the previous population of size $I_0R_0^{g-1}$, and introduce recombination and mutation events as described below.

We first tested the scenario where $\rho = 0$ and $\varepsilon = 0.003$ (we used a much larger value than the estimated 10^{-5} HIV-1 mutation rate in order to make mutations accumulate faster since by the time we reach a population size of 2^{10} or larger the simulation becomes computationally intensive). In the mutation only scenario, we randomly draw the number of positions to mutate from a Poisson distribution with mean εN_B, where N_B is the length of the sequences. If $HD_0(s, g-1)$ is the HD_0 of a sequence s at generation $g-1$, and d is randomly drawn from $Pois(\lambda = \varepsilon N_B)$, then $HD_0(s, g) = HD_0(s, g-1)+d$, and the d positions where these mutations take place are drawn randomly from a uniform distribution. Conversely, in the recombination only scenario, i.e., when $\rho = 1$ and $\varepsilon = 0$, at every generation g, for every sequence s in the sample, we draw the two parents from the previous generation and a recombination position θ uniformly and form s as the recombinant child.

In Figures 2 and 3 we show the mean HD_0 and variance HD_0 for the two scenarios respectively, averaged over 10,000 runs. The simulation is represented by the black dots, and the theoretical results are shown in the red dashed line. 95% confidence intervals are shown in gray, though in Figure 2 they are barely visible. In both scenarios the simulation results fall within the 95% confidence intervals from the theoretical expectations. Note that in Figure 3, since the fluctuations at different times are positively correlated, all points remain above the mean, following a large fluctuation in the first generation.

2.5. Mutations and Recombination

In Appendix C we generalize the methods used to derive Eq. (14) and find recursive formulae for the general case when both ρ and ε are non-zero. For the mean HD_0 we obtain

$$\mu(g)=N_B\varepsilon(1-\rho)g+\mu_0. \quad (15)$$

This shows that even in the presence of recombination, the mean HD_0 grows linearly like in the mutation only scenario. The slope of the linear growth here is diminished by a factor $1-\rho$, which reflects the fact that in our model a recombining sequence does not mutate.

Let $\xi = (1 - \rho)\varepsilon$, and let x and y be any given positions in the genome. In Appendix D we derive the following expressions for the HD_0 mean μ_x at position x, the HD_0 variance estimator s_x^2 at position x, and the HD_0 covariance estimator s_{xy} between any two positions x and y:

$$\mu_x(g+1)=\mu_x(g)+\xi \quad (16)$$

$$s_x^2(g+1)=\frac{N_g-1}{N_g}s_x^2(g)+\xi(1-\xi) \quad (17)$$

$$s_{xy}(g+1)=\left(1-\rho\frac{y-x}{N_B+1}\right)\frac{N_g-1}{N_g}s_{xy}(g)-\frac{N_{g+1}}{N_{g+1}-1}\xi^2+\xi\left(\varepsilon-\frac{1}{N_{g+1}}\right)-2\varepsilon\xi^2\frac{1-\varepsilon}{N_{g+1}} \quad (18)$$

By summing Eqs. (16), (17), and (18) across all positions x and y we get the mean HD_0 and the variance estimator $s^2(g)$ in the general case when both mutation and recombination are present. Again, we verified these theoretical derivations against numerous simulation runs. The simulation was designed as follows.

For a fixed recombination rate ρ and mutation rate ε, we first draw the number of recombination events from a binomial distribution with probability ρ. Subsequently, for each sequence s in the new generation, if s is a recombinant, we uniformly draw the two parents and a recombination position θ, and form s as the recombinant child. Else, we randomly draw a number of mutated positions from a Poisson distribution with mean εN_B, where N_B is the length of the sequences. If $HD_0(s, g - 1)$ is the HD_0 of a sequence s at generation $g - 1$, and d is randomly drawn from a Poisson distribution with mean εN_B, then $HD_0(s, g) = HD_0(s, g - 1) + d$, and the d positions where these mutations take place are drawn randomly from a uniform distribution.

Because of computational limitations, we limited the number of generations to 6–8 and simulated 10,000 sets, then averaged the HD_0 means and variances and compared the results to the theoretical results discussed in the previous sections. Instead of simulating from the very beginning of the infection, we started at generation $g = 0$ with $I_0 = 100$ sequences, each proviral sequence with an HD_0 from the founder strain randomly drawn from a Poisson distribution with mean $\lambda = 10$. The code was written in R (R Development Core Team, 2010).

In Figures 4, 5, and 6 we compare the simulation results with the theoretical derivations. The simulation results are denoted with the black dots and the theoretical derivations with the red dashed lines. Notice that because Eq. (D.3) and Eq. (D.4) were obtained up to terms of the order N^{-2}, where N is the sample size, when we carry the summation over all positions to obtain the general expression for $s^2(g)$, the fluctuations become of the order N^{-1}. In Figure 6 we omit the comparison with the theoretical derivation for the first two generations. In fact, for these, the sample size is small enough that fluctuations are sizable. However, starting from $g = 3$, theory and simulation overlap almost perfectly.

We performed a simulation with parameters taken from the early homogeneous sample SUMA, previously described in Keele et al. (2008) and Lee et al. (2009) (Figure 7). This time the initial population consisted of a single infecting strain and, at each generation, we sampled $N = 35$ sequences, which was the sample size of the original alignment. Under this particular scenario, on average, the HD variance was lower than the mean HD. This was

consistent with what observed across all homogeneous patients in Keele et al. (2008), for whom the HD variance was, on average, 4.7% lower than what expected from a Poisson model. According to our original Poisson model, SUMA was estimated to be 5 generations into the infection, which was consistent with a Fiebig stage II. In Figure 7 we show that adding recombination to the model does indeed cause the HD frequency counts to no longer follow a Poisson distribution, but the effect is so small that the estimate attained through our original model still remains valid.

3. Conclusions

We developed a model to study the rates of evolution and genetic diversification in acute HIV-1 infection. We used the model to examine the effects of recombination in acute HIV infection sampled prior to the onset of selection and during the exponential growth of the viral population. Previously, we neglected recombination and used a simplified model of infection where all infected cells produced virions and died synchronously (Lee et al., 2009). Using this model we showed that the mean intersequence HD grows linearly with time as $E(HD) \propto 2\varepsilon N_B g$, where ε is the mutation rate, N_B is the length of the sequences, and g is the generation number. We applied this model to homogeneously infected subjects (Keele et al., 2008; Wood et al., 2009) and, using the linearity in time of the mean HD, for each homogeneously infected subject, were able to estimate the time since the most recent common ancestor.

In this paper, we explored a more complex model of HIV replication. This was motivated by the fact that in general, our estimates in Keele et al. (2008) were well correlated with the time since the infection obtained from clinical data. However, in a Poisson model, we expect the mean HD_0 to be equal to the variance of the HD_0 times $\frac{N-1}{N}$, where N is the sample size. Even though not statistically divergent from a Poisson distribution, all homogeneous subjects presented in Keele et al. (2008) had a variance HD that was on average 4.7% lower than the expected theoretical value had the distribution been a Poisson. Here we showed that even in an age-dependent continuous model of infection, the mean HD_0 still grows linearly with time, and this continues to hold in a fully stochastic model. While recombination did not affect the linearity of the mean HD_0 growth, in the presence of recombination the HD_0 distribution is no longer Poisson. In this paper we modeled the new dependency of the variance HD_0 with time, as it no longer equals that of the mean HD_0. We also showed how this new framework helps explain some of the divergence between mean and variance HD in homogeneous samples previously described in Keele et al. (2008). Furthermore, because the effect is small, our work validates our previous methods that estimate the time of infection based on a Poisson model.

This work suggests that our previous, simplified model (Lee et al., 2009) provides reasonable estimates of the time and number of generations since a founder strain initiates the infection even in the presence of in vivo recombination. It is also robust against instances of in vitro recombination (Salazar-Gonzales et al., 2008), which becomes a relevant issue particularly when using 454 deep sequencing (Fischer et al., 2010). For 454 samples in particular, where sample sizes can reach the tens of thousands, our method provides a robust alternative to computationally intense methods such as those employed in the software package BEAST (Drummond et al., 2007).

Finally, we notice that the mathematical framework developed here can be used in cell division models where it is relevant to keep track of the number of divisions cells undergo (generations) (Zilman et al., 2010; Bernard et al., 2003).

Acknowledgments

This work was supported by the US Department of Energy through the LANL LDRD program, the Center for HIV/AIDS Vaccine Immunology, and NIH grants U19-AI067854-07, UM1-AI100645-01, AI028433, and OD011095.

References

Althaus CL, Bonhoeffer S. Stochastic interplay between mutation and recombination during the acuisition of drug resistance mutations in human immunodeficiency virus type 1. J Virol. 2005; 79:13572–8. [PubMed: 16227277]

Batorsky R, Kearney MF, Palmer SE, Maldarelli F, Rouzine IM, Coffin JM. Estimate of effective recombination rate and average selection coefficient for HIV in chronic infection. Proc Natl Acad Sci USA. 2011; 108:5661–6. [PubMed: 21436045]

Bernard S, Pujo-Menjouet L, Mackey MC. Analysis of cell kinetics using a cell division marker: mathematical modeling of experimental data. Biophys J. 84:3414–24. [PubMed: 12719268]

Chen J, Dang Q, Unutmaz D, Pathak VK, Maldarelli F, Powell D, Hu WS. Mechanisms of nonrandom human immunodefficiency virus type 1 infection and double infection: preference in virus entry is important but is not the sole factor. J Virol. 2005; 79:4140–49. [PubMed: 15767415]

Chen HY, DiMascio M, Perelson AS, Ho DD, Zhang L. Determination of virus burst size in vivo using a single-cycle SIV in rhesus macaques. Proc Natl Acad Sci USA. 2007; 104:19079–84. [PubMed: 18025463]

Dang Q, Chen J, Unutmaz D, Coffin JM, Pathak VK, Powell D, KewalRamani VN, Maldarelli F, Hu WS. Nonrandom HIV-1 infection and double infection via direct and cell-mediated pathways. Proc Natl Acad Sci USA. 2004; 101:632–37. [PubMed: 14707263]

De Boer RJ, Ribeiro RM, Perelson AS. Current estimates for HIV-1 production imply rapid viral clearance in lymphoid tissues. PLoS Comput Biol. 2010; 6:e1000906. [PubMed: 20824126]

Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007; 7:214–23. [PubMed: 17996036]

Fiebig EW, Wright DJ, Rawal BD, Garrett PE, Schumacher RT, et al. Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection. AIDS. 2003; 17:1871–79. [PubMed: 12960819]

Fischer W, Bhattacharya T, Keele BF, Giorgi EE, Hraber PT, Perelson AS, Shaw GM, Korber BT, et al. Rapid mutational escape from cytotoxic T-cell responses in acute HIV-1 infection—an ultra-deep view. PLoS ONE. 2010; 5:e12303. [PubMed: 20808830]

Fraser C. HIV recombination: what is the impact on antiretroviral therapy? J R Soc Interface. 2005; 2(489)

Gelfand, Fomin SV. Calculus of Variations. Englewood Cliffs, NJ: 1963.

Giorgi EE, Funkhouser B, Athreya G, Perelson AS, Korber BT, Bhattacharya T. Estimating time since infection in early homogeneous HIV-1 samples using a Poisson model. BMC Bioinformatics. 2010; 11:532–9. [PubMed: 20973976]

Gillespie JH. The status of the neutral theory. Science. 1984; 224:732–33. [PubMed: 17780612]

Huddleston JV. Population Dynamics with age- and time-dependent birth and death rates. Bull of Math Biol. 1983; 45:827–36.

Jetzt AE, Yu H, Klarmann GJ, Ron Y, Preston BD, Dougherty JP. High rate of recombination throughout the human immunodeficiency virus type I genome. J Virol. 2000; 74:1234–40. [PubMed: 10627533]

Jung A, Meier R, Vartanian JP, Bocharov G, Jung V, Fischer U, Meese E, Wain-Hobson S, Meyerhans A. Multiply infected spleen cells in HIV patients. Nature. 2002; 418:144. [PubMed: 12110879]

Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, et al. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proc Natl Acad Sci USA. 2008; 105:7552–57. [PubMed: 18490657]

Keele BF, Li H, Learn GH, Hraber P, Giorgi EE, Grayson T, Sun C, Chen Y, Yeh WW, Letvin NL, Mascola JR, Nabel GJ, Haynes BF, Bhattacharya T, Perelson AS, Korber BT, Hahn BH, Shaw

GM. Low-dose rectal inoculation of rhesus macaques by SIVsmE660 or SIVmac251 recapitulates human mucosal infection by HIV-1. J Exp Med. 2009; 206:1117–34. [PubMed: 19414559]

Kevorkian, J. Partial Differential Equations: Analytical Solutions Techniques. 2. Springer-Verlag; New York: 2002.

Kimura M. Evolutionary rate at the molecular level. Nature. 1968; 217:624–626. [PubMed: 5637732]

Kouyos RD, Fouchet D, Bonhoffer S. Recombination and drug resistance in HIV: population dynamics and stochasticity. Epidemics. 2009; 1:58–69. [PubMed: 21352751]

Lee HY, Giorgi EE, et al. Modeling sequence evolution in acute HIV-1 infection. J Theor Biol. 2009; 261:341–60. [PubMed: 19660475]

Levy DN, Aldrovandi GM, Kutsch O, Shaw GM. Dynamics of HIV-1 recombination in its natural target cells. Proc Natl Acad Sci USA. 2004; 101:4204–09. [PubMed: 15010526]

Markowitz M, Louie M, Hurley A, Sun E, Di Mascio M, et al. A novel antiviral intervention results in more accurate assessment of human immunodeficiency virus type 1 replication dynamics and T-cell decay in vivo. J Virol. 2003; 77:5037–38. [PubMed: 12663814]

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005; 437:376–380. [PubMed: 16056220]

Mostowy R, Kouyos RD, Fouchet D, Bonhoffer S. The role of recombination for the coevolutionary dynamics of HIV and the immune response. PLoS One. 2011; 6:e16052. [PubMed: 21364750]

Moutouh L, Corbeil J, Richman DD. Recombination leads to the rapid emergency of HIV-1 dually resistant mutants under selective drug pressure. Proc Nat Acad Sci USA. 1996; 93:6106–11. [PubMed: 8650227]

Murray, JD. Mathematical Biology. 3. Springer-Verlag; Berlin Heidelberg: 2002.

Neher RA, Leitner T. Recombination rate and selection strength in HIV intra-patient evolution. PLoS Comput Biol. 2010; 6:e1000660. [PubMed: 20126527]

Nisbet, RM.; Gurney, WSC. Modeling Fluctuating Populations. J Wiley and Sons; 1982.

Onafuwa-Nuga A, Telesnitsky A. The remarkable frequency of human immunodeficiency virus type 1 genetic recombination. Microbiol Mol Biol Rev. 2009; 73:451–80. [PubMed: 19721086]

Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science. 1996; 271:1582–1586. [PubMed: 8599114]

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2010. URL http://www.R-project.org

Ramirez BC, Simon-Loriere E, Galetto R, Negroni M. Implications of recombination for HIV diversity. Vir Res. 2008; 134:64–73.

Ramratnam B, Bonhoeffer S, Binley J, Hurley A, Zhang L, et al. Rapid production and clearance of HIV-1 and hepatitis C virus assessed by large volume plasma apheresis. Lancet. 1999; 354:1782–85. [PubMed: 10577640]

Reichl, LE. A Modern Course in Statistical Physics. John Wiley and Sons; 1998.

Robertson DL, Sharp PM, McCutchan FE, Hahn BH. Recombination in HIV-1. Nature. 1995; 374:124. [PubMed: 7877682]

Ribeiro RM, Qin L, Chavez LL, Li D, Self SG, Perelson AS. Estimation of the initial viral growth rate and basic reproductive number during acute HIV-1 infection. J Virol. 2010; 12:6096–102. [PubMed: 20357090]

Sabino EC, Shpaer EG, Morgado MG, Korber BT, Diaz RS, Bongertz V, Cavalcante S, Galvo-Castro B, Mullins JI, Mayer A. Identification of human immunodeficiency virus type 1 envelope genes recombinant between subtypes B and F in two epidemiologically linked individuals from Brazil. J Virol. 1994; 68:6340–6. [PubMed: 8083973]

Salazar-Gonzalez JF, Bailes E, Pham KT, Salazar MG, Guffey MB, Keele BF, Derdeyn CA, Farmer P, Hunter E, Allen S, Manigart O, Mulenga J, Anderson JA, Swanstrom R, Haynes BF, Athreya GS, Korber BT, Sharp PM, Shaw GM, Hahn BH. Deciphering human immunodefficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. J Virol. 2008; 82:3952–70. [PubMed: 18256145]

Shriner D, Rodrigo AG, Nickle DC, Mullins JI. Pervasive genomic recombination of HIV-1 in vivo. Genetics. 2004; 167:1573–83. [PubMed: 15342499]

Suryavanshi GW, Dixit NM. Emergence of recombinant forms of HIV: dynamics and scaling. PLoS Comp Biol. 2007; 3:e205.

Vijay NNV, Ajmani R, Perelson AS, Dixit NM. Recombination increases human immunodeficiency virus fitness, but not necessarily diversity. J Gen Virology. 2008; 89:1467. [PubMed: 18474563]

Von Foerster, H. Some remarks on changing populations. In: Stohlman, F., Jr, editor. The Kinetic of Cellular Proliferation. Grune & Stratton; New York: 1959. p. 382-407.

Wallstrom, TC.; Daniels, MG.; Bhattacharya, T. Personal communication. 2010.

Wood N, Bhattacharya T, Keele BF, Giorgi EE, Liu M, Gaschen B, Daniels M, Ferrari G, Haynes BF, McMichael A, Shaw GM, Hahn BH, Korber B, Seoighe C. HIV evolution in early infection: selection pressures, patterns of insertion and deletion, and the impact of APOBEC. PLoS Pathog. 2009; 5:e1000414. [PubMed: 19424423]

Wooley DP, Smith RA, Czajak S, Desrosiers RC. Direct demonstration of retroviral recombination in rhesus monkey. J Virol. 1997; 71:9650–3. [PubMed: 9371629]

Zhang M, Foley B, Schultz AK, Macke JP, Bulla I, Stanke M, Morgenstern B, Korber B, Leitner T. The role of recombination in the emergence of a complex and dynamic HIV epidemic. Retrovirology. 2010; 7:25. [PubMed: 20331894]

Zilman A, Ganusov VV, Perelson AS. Stochastic models of lymphocyte proliferation and death. PLoS ONE. 2010; 5:e12775. [PubMed: 20941358]

Appendix A. Stochastic model

We present a stochastic formulation of the continuous model presented in Section 2.1. In all that follow we use the same mathematical framework developed for birth and death stochastic processes, with the added parameter g for generation. To keep the discussion more general, we will use the word “births” instead of “new infections.” Clearly, when applied to the infection model described in this paper, new births are newly infected cells.

Denote $U = \mathbb{R}_+ \times \mathbb{Z}_+$ the space of all pairs $u = (a, g)$, as defined in Section 2. We define a state of the system as a function $I: U \to \mathbb{Z}_+$ that associates to every $u \in U$ a positive integer $I(u) = I(a, g)$, the number of infected cells of age a and generation g.

Let $\mathcal{W}$ be the space of all functions $I: \mathbb{R}_+ \times \mathbb{Z}_+ \to \mathbb{Z}_+$. At any given time $t \in \mathbb{R}_+$, and for any given state $I \in \mathcal{W}$, let $P(I; t)$ be the probability of the system being described by the state I at time t. Even though $P: \mathcal{W} \times \mathbb{R}_+ \to [0, 1]$, for simplicity we omit the dependency in time and consider P as an element in $S(\mathcal{W})$, where $S(\mathcal{W}) = \{\mathcal{W} \to [0, 1]\}$, the space of all operators from $\mathcal{W}$ into [0, 1].

To better illustrate the derivation, we first show how to obtain the master equation discretized in time increments Δt. Define $\Delta: \mathbb{R}_+ \times \mathcal{W} \to \mathcal{W}$ as follows. Given $\tau \in \mathbb{R}_+$, $I \in \mathcal{W}$, and $(a, g) \in U$, then $\Delta_\tau I(a, g) = I(a + \tau, g) - I(a, g)$. From the above notice that for every $I \in \mathcal{W}$ and every $(a, g) \in U$, $(I + \Delta_\tau I)(a, g) = I(a + \tau, g)$. $\Delta_\tau I$ represents the difference between state I at time $t + \tau$ and state I at time t when cells age only (e.g. no deaths nor births). It follows that, given $I_{t+\tau}$, the state at time t is $I_{t+\tau} - \Delta_\tau I$.

We also define the Kroenecker Delta operator as $\delta: U \to \mathcal{W}$ as follows. Given $u_0 \in U$, $\delta_{u_0} \in \mathcal{W}$ is defined so that for every $u \in U$ such that $u \neq u_0$, $\delta_{u_0}(u) = 0$ and $\delta_{u_0}(u_0) = 1$.

Suppose now that at a given time t, the system is completely described by the state I. We wish to compute the probability $P(I, t + \Delta t)$. In other words, we want to see how the system evolves after an increment in time of Δt. There are three possible contributions to the term $P(I, t + \Delta t)$. In what follows we assume that for every age a and generation g, out of all infected cells $I(a, g)$ only one can die or be born at the time. Hence, the contributions are:

1. No deaths, no births, cells can only survive. The contribution in this case is

$$P(I-\Delta_{\Delta t}I;t)\prod_{u\in U}(1-\beta(u)\Delta t)^{(I-\Delta_{\Delta t}I)(u)}\prod_{u\in U}(1-\alpha(u)\Delta t)^{(I-\Delta_{\Delta t}I)(u)} \quad \text{(A.1)}$$

2. One cell u_0 dies. The contribution (summed over all possible $u_0 \in U$) is

$$\sum_{u_0\in U}P(I+\delta_{u_0};t)\left(\prod_{u\neq u_0}(1-\beta(u)\Delta t)^{I(u)}\right)\beta(u_0)\Delta t\,(I(u_0)+1) \quad \text{(A.2)}$$

3. One cell u_0 is born. The contribution (summed over all possible $u_0 \in U$) is

$$\sum_{u\in U}P(I-\delta_{u_0=(0,g+1)};t)\left(\prod_u(1-\beta(u)\Delta t)^{I(u)}\right)\alpha(u)\Delta t\left(I(u)-\delta_{u_0=(0,g+1)}\right) \quad \text{(A.3)}$$

We expand each of the above terms in Δt and neglect terms in $O(\Delta t^2)$ to obtain, for each piece respectively:

1.
$$P(I-\Delta_{\Delta t}I;t)\left(1-\sum_{u\in U}(I-\Delta_{\Delta t}I)(u)\,[\beta(u)+\alpha(u)]\,\Delta t\right) \quad \text{(A.4)}$$

2.
$$\sum_{u\in U}P(I+\delta_u;t)\,(I(u)+1)\;\beta(u)\Delta t \quad \text{(A.5)}$$

3.
$$\sum_{u\in U}P(I-\delta_{u_0=(0,g+1)};t)\,(I(u)-\delta_{u_0=(0,g+1)})\;\alpha(u)\Delta t \quad \text{(A.6)}$$

The first order master equation of the process, discretized in time, is therefore

$$\begin{aligned}&\tfrac{P(I;t+\Delta t)-P(I-\Delta_{\Delta t}I;t)}{\Delta t}=\\&=-P(I-\Delta_{\Delta t}I;t)\,(\textstyle\sum_{u\in U}(I-\Delta_{\Delta t}I)(u)\,[\beta(u)+\alpha(u)])+\\&+\textstyle\sum_{u\in U}P(I+\delta_u;t)\,(I(u)+1)\;\beta(u)+\\&+\textstyle\sum_{u\in U}P(I-\delta_{(0,g+1)};t)\left(I(u)-\delta_{u_0=(0,g+1)}\right)\;\alpha(u)\end{aligned} \quad \text{(A.7)}$$

To obtain the continuous master equation we have to take the limit for $\Delta t \to 0$ on both sides of Eq. (A.7). In order to do so, we use an approximation that is a direct consequence of the definition of functional derivative (Gelfand et al., 1963): given $x, y \in \mathcal{W}$ and $f \in \mathcal{S}(\mathcal{W})$, the following holds

$$f(x+y) \simeq f(x)+\int_U y(u)\frac{\partial f}{\partial x(u)}du \quad \text{(A.8)}$$

Using the above, the left-hand side of Eq. (A.7) becomes

$$\frac{\partial P}{\partial t} - \int_U \frac{\partial I}{\partial a}(u) \frac{\partial P}{\partial I(u)} du \quad \text{(A.9)}$$

Notice that in the absence of births and deaths (i.e. $\alpha = \beta = 0$), the above reduces to

$$\frac{\partial P}{\partial t}(I;t) - \int_U \frac{\partial I}{\partial a}(u) \ \frac{\partial P}{\partial I(u)}(I;t)\, du = 0 \quad \text{(A.10)}$$

which describes the survival of all cells present at time $t = 0$.

In order to expand the right-hand side (RHS) of Eq. (A.7), we collect the terms in β together, and similarly for the terms in α. We approximate to the first order in Δt and let $\Delta t \to 0$.

$$\text{RHS}(\beta) = P(I;t)\left(\sum_u I(u)\beta(u) + \sum_u \beta(u) - \sum_u I(u)\beta(u)\right) + \sum_u \left(\frac{\partial P}{\partial I(u)}(I)\right) I(u)\ \beta(u) + o(\Delta t) \quad \text{(A.11)}$$

In the limit the above becomes

$$P(I;t)\int_U \beta(u) du + \int_U \beta(u) \frac{\partial P}{\partial I(u)} I(u)\, du \quad \text{(A.12)}$$

By expanding RHS(α) in a similar way, one obtains the continuous master equation

$$\begin{aligned} \frac{\partial P}{\partial t}(I;t) &= \int_U \frac{\partial I}{\partial a}(u) \ \frac{\partial P}{\partial I(u)}(I;t) \ du = \\ &= \int_U \mu(u) \left[P(I;t) + \frac{\partial P}{\partial I(u)}(I;t) \ I(u)\right] du + \\ &- \int_{\partial U} \alpha(u) \left[P(I;t) - \frac{\partial P}{\partial I(u)}(I;t) \ I(u)\right] du + \\ &- \int_U \frac{\partial P}{\partial I(0,g+1)}(I;t) \ I(u) \ \alpha(u) \ du \end{aligned} \quad \text{(A.13)}$$

where $\partial U = \{u = (a, g) \in U | a = 0, g > 0\}$. In all that follows we assume that newborn cells cannot give birth, which implies that $\alpha_{|\partial U} = 0$. As a consequence, the second term on the RHS is null. We now use the following two claims to re-write the above equation.

Claim Appendix A.1

Let $u \in U$. We define a map $S(\mathcal{W}) \to S(\mathcal{W})$ that to every $P \in S(\mathcal{W})$ associates a $\tilde{P}_u \in S(\mathcal{W})$ so defined: $\tilde{P}_u(I) = I(u)P(I)$. Then

$$\frac{\partial \tilde{P}_u(I)}{\partial I(u)} = P(I) + I(u)\frac{\partial P(I)}{\partial I(u)} \quad \text{(A.14)}$$

Proof

Given $\varepsilon > 0$ and $u_0 \in U$, define $\delta_{u_0,\varepsilon} \in \mathcal{W}$ as the following Kroencker delta: $\delta_{u_0,\varepsilon}(u) = \varepsilon\, \delta(u - u_0)$. Then

$$\tilde{P}_u(I + \delta_{u,\varepsilon}) - \tilde{P}_u(I) = (I + \delta_{u,\varepsilon})(u)P(I + \delta_{u_0,\varepsilon}) - I(u)P(I) = \delta_{u,\varepsilon}(u)P(I + \delta_{u_0,\varepsilon}) + I(u)\left[P(I + \delta_{u,\varepsilon}) - P(I)\right] \quad \text{(A.15)}$$

By dividing by ε and letting $\varepsilon \to 0$ we get the assert.

Claim Appendix A.2

Given $u_0 \neq u$, we have

$$\frac{\partial \tilde{P}_u(I)}{\partial I(u_0)} = I(u) \frac{\partial P(I)}{\partial I(u_0)} \quad \text{(A.16)}$$

As a consequence, we can rewrite the master equation as follows

$$\begin{aligned} &\frac{\partial P}{\partial t}(I;t) - \int_U \frac{\partial I}{\partial a}(u) \ \frac{\partial P}{\partial I(u)}(I;t) \ du \\ &= \int_U \frac{\partial}{\partial I(u)} [I(u)P(I;t)] \ \mu(u) \ du \\ &- \int_U \frac{\partial}{\partial I(u_0)} \ [I(u)P(I;t)] \ \alpha(u) \ du \end{aligned} \quad \text{(A.17)}$$

where $u_0 = (0, g + 1)$.

Eq. (A.17) reduces to well known results when the birth and death rates are both constant, in which case, if we integrate both sides over the domain $\mathcal{W}_N = \{I \in \mathcal{W} | \int_U I = N_t\}$, where N_t is the total number of cells at time t, and N_0 is the initial number of cells. we obtain

$$\frac{\partial P}{\partial t}(N_t;t) = \beta [(N_t+1) P(N_t+1;t) - N_t P(N_t;t)] + \alpha [(N_t-1) P(N_t-1;t) - N_t P(N_t;t)] \quad \text{(A.18)}$$

which is the master equation of the classic linear birth and death process. This is fully discussed in Reichl (1998), where a solution is given in terms of the generating function $F(z; t) = \Sigma_{N_t} P(N_t; t) \ z^{N_t}$ by

$$F(z;t) = \left[\frac{\beta(z-1)e^{(\alpha-\beta)t} - \alpha \ z + \beta}{\alpha(z-1)e^{(\alpha-\beta)t} - \alpha \ z + \beta} \right]^{N_0}. \quad \text{(A.19)}$$

Here N_0 is the initial number of infected cells.

Appendix B. Continuous model, closed form solution

We now develop the closed form solution to Eq. (1). One can verify that, given a smooth function $\varepsilon(g, x)$ such that $\varepsilon(g, x) = 0$ when $x < 0$, then

$$I(a, g, t) = \phi(g, t-a) \ e^{-\int_0^a \beta(s) ds} \quad \text{(B.1)}$$

satisfies the first equation in (1). Substituting into the renewal condition, one gets that ε should also satisfy

$$\phi(g+1, t) = \int_\Lambda^\infty \alpha(a) \ e^{-\int_0^a \beta(s) ds} \phi(g, t-a) \, da. \quad \text{(B.2)}$$

Given the initial population of I_0 cells, at any given time t, the only cells at $g = 0$ are the survivors from the initial population. In other words:

$$I(a,0,t)=\begin{cases} I_0 e^{-\int_0^a \beta(s)ds} & \text{if} \quad a=t \\ 0 & \text{otherwise.} \end{cases} \quad \text{(B.3)}$$

It follows that for $g = 0$, ε is completely determined by $\varepsilon(0, t - a) = \delta(t - a)I_0$, where $\delta(t - a)$ is the singular density. When $g = 1$

$$I_0 \int_0^\infty f(a)\delta(a-t)\ da = I_0 f(t) = \phi(1,t) \quad \text{(B.4)}$$

and therefore one can show recursively that for $g \geq 1$

$$\phi(g,t) = I_0 \int_0^\infty \cdots \int_0^\infty \left(\prod_{i=1}^{g} f(a_i)\, da_i \right) \delta\left(\sum_{i=1}^{g} a_i - t \right). \quad \text{(B.5)}$$

By reordering the above and substituting f with its Fourier transform $\tilde{f}$ one obtains

$$\begin{aligned} \phi(g,t) &= I_0 \int_0^\infty \cdots \int_0^\infty \int_{-\infty}^\infty e^{ik(\sum a_i - t)} \frac{dk}{2\pi} \left(\prod_{i=1}^{g} f(a_i)\ da_i \right) = \\ &= \frac{I_0}{2\pi} \int_{-\infty}^\infty e^{-ikt} \left(\prod_{i=1}^{g} \int_0^\infty f(a_i)\, e^{ika_i}\, da_i \right)\ dk = \\ &= \frac{I_0}{2\pi} \int_{-\infty}^\infty e^{-ikt}\, \tilde{f}(k)^g\, dk \end{aligned} \quad \text{(B.6)}$$

Substituting Eq. (B.6) in Eq. (B.1), one gets Eq. (4):

$$I(g,t) = \frac{I_0}{2\pi} \int_0^t e^{-\int_0^a \beta(s)ds} \int_{-\infty}^\infty \tilde{f}(k)^g\, e^{-ik(t-a)}\, dk\, da \quad \text{(B.7)}$$

which reduces the problem to a quadrature.

Appendix C. HD variance for an exponentially growing population in the presence of recombination alone

We now illustrate how to obtain expressions for the mean HD_0 and the expected value of the variance of HD_0 as functions of generation g in the case where the population grows exponentially, the mutation rate is $\varepsilon = 0$, and the recombination rate is $\rho = 1$.

For every position x between 1 and N_B, we compute both the mean HD_0 and the HD_0 variance at that position across all sequences, and for every pair of positions x and y, we will compute the covariance across all sequences, and finally we will obtain the total mean HD_0 and HD_0 variance by summing across all positions.

Since there are no new mutations, at any generation g, the mean HD at d position x is given by $\overline{HD}_x(g) = \frac{d}{N_B}$, and hence

$$\mu(g) = \sum_{x=1}^{N_B} \overline{HD}_x(g) = d. \quad \text{(C.1)}$$

An unbiased estimator of the variance at position x is given by

$$s_x^2(g)=\frac{N_g}{N_g-1}\left(\overline{HD_x^2}-\overline{HD_x}^2\right), \quad \text{(C.2)}$$

Here, for simplicity of notation, we have omitted the dependency in g. So, if $s_x^2(g+1)$ is the expected variance at generation $g+1$, summing over all sequences s we obtain

$$s^2(g+1)=\frac{1}{N_g}\left(\sum_s\left(HD_x-\frac{1}{N_g}\sum_{s'}HD_x\right)^2\right)=\frac{N_g-1}{N_g}s^2(g). \quad \text{(C.3)}$$

This is not quite the total variance as we still need to add in the contribution from the covariance between every pair of positions x and y.

To understand how the expected covariance $s_{x,y}$ between any two sites x and y changes with generation g, we make the following observation: if the split position θ falls between the two positions, then x and y will be inherited independently to the next generation. If, on the other hand, θ does *not* split the two positions, then the new covariance will be a random variate with expectation the second cumulant of the covariance at the previous distribution. In order to use this observation, we split the set of sequences in two groups: the ones derived from two parent sequences which recombined somewhere between x and y, and those derived instead from a recombination split *outside* x and y.

Let N_1 be the number of sequences in the first group, and N_2 the number of sequences in the second group, so that $N_1 + N_2 = N_g$ (when clear from the context, we omit the dependency in g for simplicity of notation). We define $w_i=E\left(\frac{N_i}{N_g}\right)$ and κ_{1x}^i and κ_{11}^i the cumulants of the HD in the i-th group at a fixed generation g. An unbiased estimator of the covariance is given by

$$s_{xy}(g)=\frac{1}{N_g-1}\sum_{s\in N_g}(HD_x-\overline{HD}_x)(HD_y-\overline{HD}_y) \quad \text{(C.4)}$$

where N_g is the total number of sequences at generation g, HD_x and HD_y are the Hamming distances at positions x and y respectively, and $\overline{HD}_x$ and $\overline{HD}_y$ their means. By splitting the summation over sequences in N_1 and sequences in N_2, one gets

$$s_{xy}(g)=\sum_{i=1,2}\frac{N_i-1}{N_g-1}s_i+\sum_{i=1,2}\frac{N_i}{N_g-1}(\overline{HD}_{x,i}-\overline{HD}_x)(\overline{HD}_{y,i}-\overline{HD}_y) \quad \text{(C.5)}$$

where s_i is the variance over subset N_i, and $\overline{HD}_{x,i}$ and $\overline{HD}_{y,i}$ are the means at positions x and y respectively within group N_i. Taking expectations on both sides of the above equation, one obtains

$$E(s_{xy})|_{g+1}=\sum_{i=1,2}w_i\kappa_{11}^i+\frac{N_{g+1}}{N_{g+1}-1}\left[\sum_{i=1,2}w_i\kappa_{1x}^i\kappa_{1y}^i-\left(\sum_{i=1,2}w_i\kappa_{1x}^i\right)\left(\sum_{j=1,2}w_j\kappa_{1y}^j\right)\right]. \quad \text{(C.6)}$$

Notice that the first term on the right hand side is the weighted average of the κ_{11}^{i}'s, and the second one is the weighted covariance between κ_{1x}^{i} and κ_{1y}^{i}.

The first cumulants are the mean of the corresponding distribution, hence $\kappa_{1x}^{i}=\kappa_{1y}^{i}=d$ at all times, whereas with the above definitions of N_1 and N_2, we get $\kappa_{11}^{1}=0$ (the split position separates x and y and hence the number of mutations at each site become independent variables), and $\kappa_{11}^{2}=\frac{N_g-1}{N_g}E(s_{xy})$. Furthermore, $w_2=\left(1-\frac{y-x}{N_B+1}\right)$. Therefore, substituting all of the above into Equation (C.6), we obtain

$$E(s_{xy})(g+1)=\left(1-\frac{y-x}{N_B+1}\right)\frac{N_g-1}{N_g}E(s_{xy})(g) \quad \text{(C.7)}$$

By adding across all positions we get

$$s^2(g)=\sum_{x=1}^{N_B}s_x^2(g)+2\sum_{x<y}s_{xy}(g)= \\ =\prod_{j=0}^{g-1}\frac{N_j-1}{N_j}\left[\sum_{x=1}^{N_B}s_x^2(0)+2\sum_{x<y}\left(1-\frac{y-x}{N_B+1}\right)^{g}s_{x,y}(0)\right] \quad \text{(C.8)}$$

With these recursive relations, we have reduced the problem to computing the initial quantities $s_x^2(0)$ and $s_{x,y}(0)$. Consider the initial population to be made of I_0 sequences such that the HD_0 of any given sequence s from the founder strain s_0 is a random variable from a probability distribution P with mean and variance μ_0 and σ_0 respectively. Then the moment generating function M of the HD_0 distribution is given by

$$M(u_1,\ldots,u_{N_B-1})=\left[\frac{1}{N_B}\left(1+\sum_{i=1}^{N_B-1}e^{u_i}\right)\right]^{\mu_0 N_B^{-1}}. \quad \text{(C.9)}$$

Here N_B is the length of the genomes, and $u_1, \ldots u_{N_B-1}$ are independent variables representing the positions in the genome (there's a total of N_B positions, but only $N_B - 1$ are independent due to the constraint that the initial mean is μ_0). Taking the first and second derivatives, we find

$$s_x^2(0)=\frac{N_B-1}{N_B^2}\mu_0+\frac{1}{N_B^2}\sigma_0^2 \quad \text{(C.10)}$$

and

$$s_{xy}(0)=\frac{1}{N_B^2}(\sigma_0^2-\mu_0) \quad \text{(C.11)}$$

It's easy to verify that summing the above two equations over all positions, one gets the initial variance σ_0^2.

Combining equations (C.8), (C.10), and (C.11), we get the expression for the total HD variance at generation $g > 0$:

$$\sigma_\rho^2(g)=d\left(1-\frac{1}{N_B}\right)-2\frac{d}{N_B^2}\sum_{x<y\le N_B}\left(1-\frac{y-x}{N_B+1}\right)^g\prod_{j=0}^{g-1}\frac{N_j-1}{N_j}=$$
$$=d\left(1-\frac{1}{N_B}\right)-2\frac{d}{g+2}\left[\left(\frac{N_B}{N_B+1}\right)^{g-1}+O(N_B^{-1})\right]\prod_{j=0}^{g-1}\frac{N_j-1}{N_j} \quad \text{(C.12)}$$

where we have used the following relationship:

$$2\sum_{x=1}^{N_B-1}\sum_{y=x+1}^{N_B}\left(1-\frac{y-x}{N_B+1}\right)^g=\frac{2}{(N_B+1)^g}\sum_{k=1}^{N_B-1}k(k+1)^g=\frac{N_B^{g+1}}{(g+2)(N_B+1)^{g-1}}+O(N_B). \quad \text{(C.13)}$$

Appendix D. HD variance for an exponentially growing population when both recombination and mutation rate are non-null

We now extend the methods used to derive Eq. (14) and find recursive formulae for the general case when both ρ and ε are non-null. Eq. (C.6) can be generalized to an indefinite number of partitions N_i such that $\Sigma_i N_i = N_g$, the total number of sequences at generation g. We use this to find analogous recursive formulae for the cumulants of the mean HD_0 distribution when both recombination and mutation events are present.

Fix positions x and y and choose N_1 to be the number of sequences derived from a recombination event that split positions x and y; N_2 the number of sequences derived from either a recombination event that did *not* split positions x and y, or from a division with no mutation; and, finally, denote N_k, with $k = 3$, 4 or 5 the number of sequences such that only the x position has mutated, only the y position, or both, respectively. Like before, let $w_i=E\left(\frac{N_i}{N_g}\right)$. Assume that mutations happen with a rate ε (per site, per generation), and recombination events happen with a rate ρ. Given this set-up, we notice that

$$E(w_j)=\begin{cases}\frac{y-x}{N_B+1}\rho & \text{if} \quad \text{j=1}\\ \left(1-\frac{y-x}{N_B+1}\right)\rho+(1-\rho)(1-\varepsilon) & \text{if} \quad \text{j=2}\\ \varepsilon(1-\rho)(1-\varepsilon) & \text{if} \quad \text{j=3 or 4}\\ (1-\rho)\varepsilon^2 & \text{otherwise.}\end{cases} \quad \text{(D.1)}$$

Let $\xi = \varepsilon(1 - \rho)$. Carrying out similar computations as sketched in the previous section, one notices that, for every positions x and y

$$\mu_x(g+1)=\mu_x(g)+\xi \quad \text{(D.2)}$$

$$s_x^2(g+1)=\frac{N_g-1}{N_g}s_x^2(g)+\xi(1-\xi) \quad \text{(D.3)}$$

$$s_{xy}(g+1)=\left(1-\rho\frac{y-x}{N_B+1}\right)\frac{N_g-1}{N_g}s_{xy}(g)-\frac{N_{g+1}}{N_{g+1}-1}\xi^2+\xi\left(\varepsilon-\frac{1}{N_{g+1}}\right)-2\varepsilon\xi^2\frac{1-\varepsilon}{N_{g+1}} \quad \text{(D.4)}$$

By summing Eq. (D.2) over all positions x one obtains Eq. (15). Similarly, by summing Eq. (D.3) and Eq. (D.4) across all positions, and noticing that we can still use the same initial values we used in Appendix C, one obtains the general expression for $s^2(g)$.

Highlights

- We describe a continuous time, age-dependent model of early HIV infection.
- We show that the new model yields the same results as a simplified model we published in 2009.
- We introduce recombination in the model.
- We show how recombination affects the results compared to our previous model.

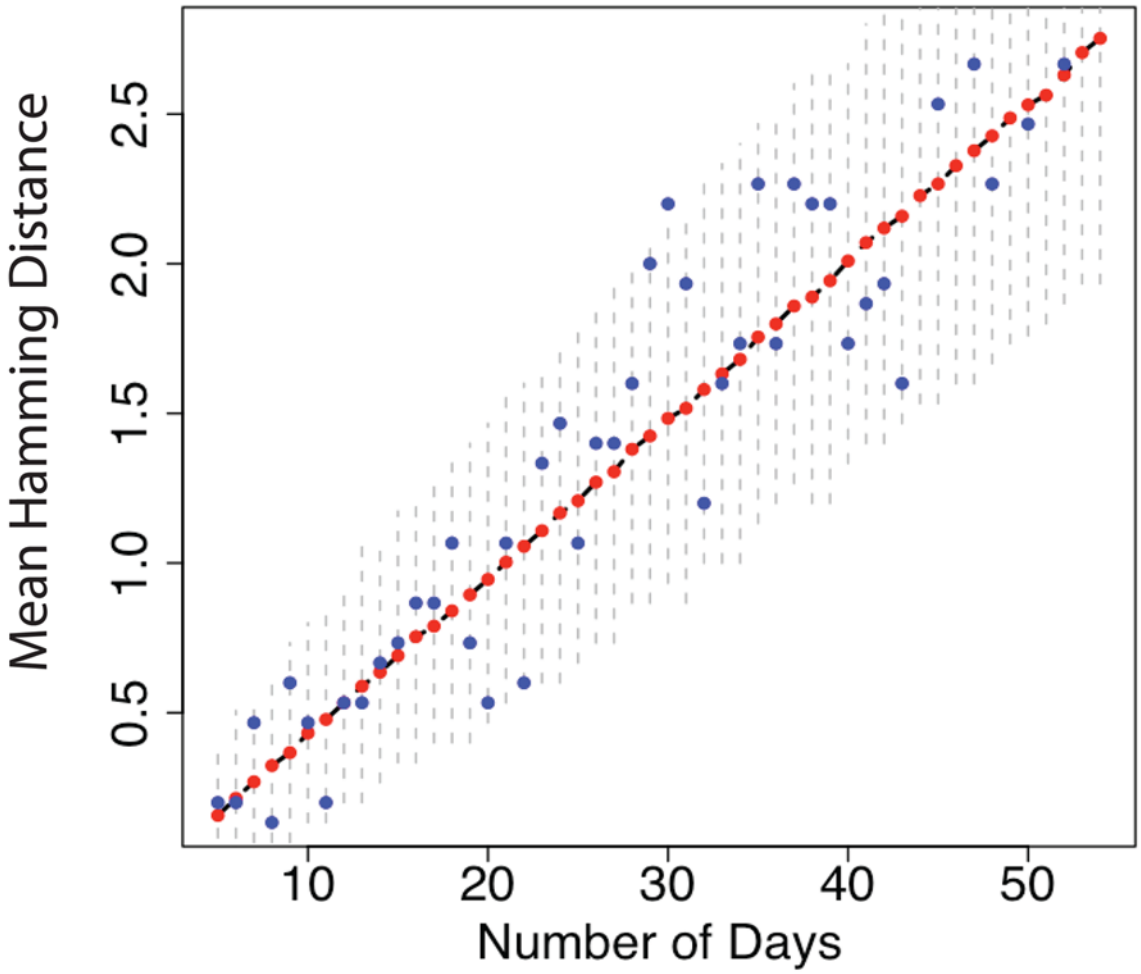

Figure 1.

We generated $I(g, t)$ using the continuous, age-dependent model described in Section 1. Here $N_B = 2,600$ base pairs, $\alpha(a) = \alpha_0(a - \Lambda)$ for $a > \Lambda$, and $\beta(a) = \beta_0 a$. At each day we sampled $N_0 = 30$ sequences, calculated the mean HD and then averaged over 10,000 runs. The blue dots shot the mean HD for one particular run, the red dots the average over all runs, and the black dashed line underneath shows the mean HD calculated from the theoretical framework. The vertical dashed lines are the 95% confidence intervals.

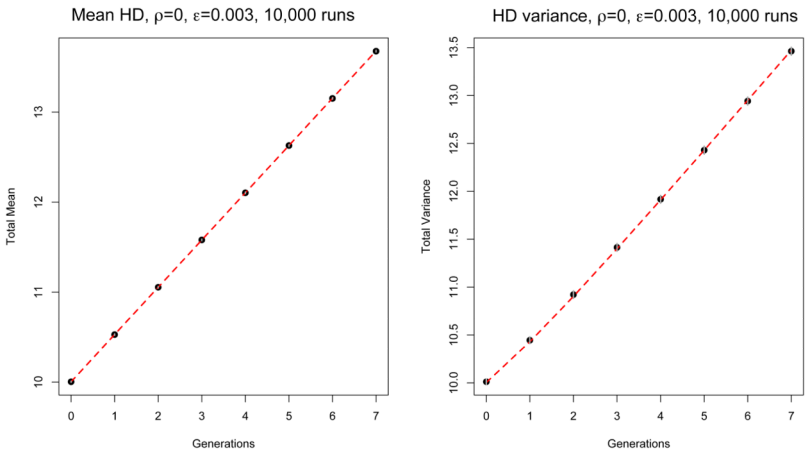

Figure 2.

Comparison between theory (red dashed line) and simulation (black dots) for an exponentially growing population with $\rho = 0$ (no recombination) and $\varepsilon = 0.003$. Mean and variance of the HDs are averaged over 10,000 runs. At generation $g = 0$ the initial population is made of 100 sequences with HD mean and variance $\lambda_0 = 10$. Confidence intervals (±1.96 times the standard deviation over runs) are represented by vertical gray segments, which are smaller than the size of the symbols in the figure and hence are not visible.

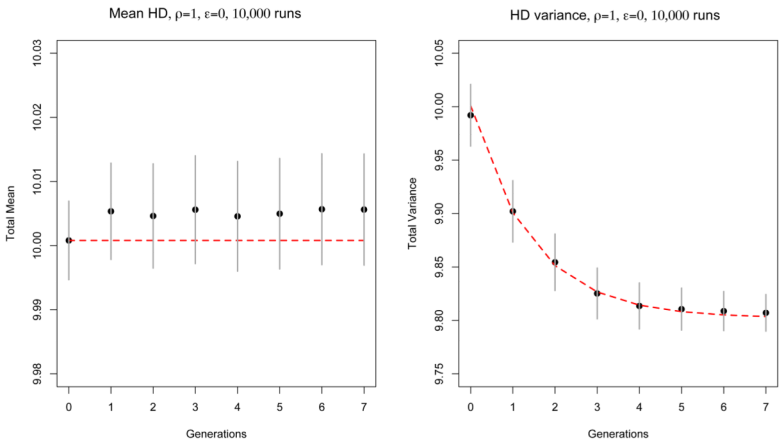

Figure 3.

Comparison between theory (red dashed line) and simulation (black dots) for an exponentially growing population with $\rho = 1$ and $\varepsilon = 0$ (recombination only). Mean and variance of HD are averaged over 10,000 runs. At generation $g = 0$ the initial population is made of 100 sequences with HD mean and variance $\lambda_0 = 10$. Confidence intervals (±1.96 times the standard deviation over runs) are represented by vertical gray segments. Errors at different time points are positively correlated.

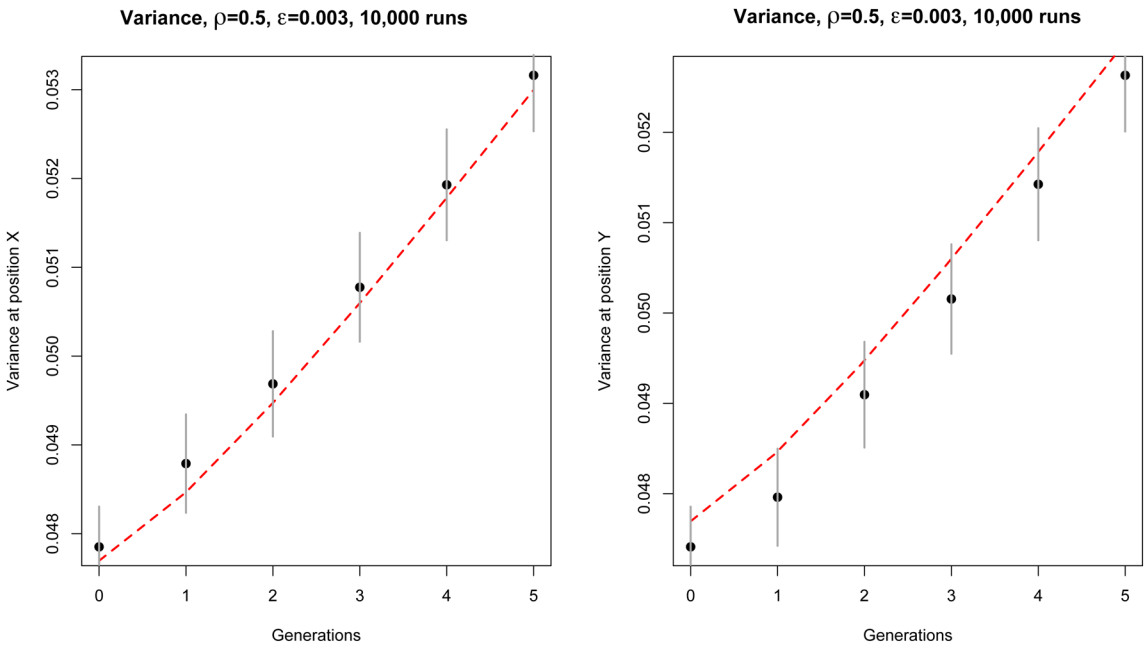

Figure 4.

Comparison between theory (red dashed line) and simulation (black dots) of the contribution of each position (denoted x and y, located at one third and two thirds of the genome) to the variance of HD_0 for an exponentially growing population. Here $\rho = 0.5$ and $\varepsilon = 0.003$. These results were obtained from 10,000 independent simulation runs with the same initial conditions, as described in the main text. The simulation is represented by the black dots, and the theory by the dashed red line. Confidence intervals (±1.96 times the standard deviation over runs) are represented by vertical gray segments. Notice that the errors are positively correlated.

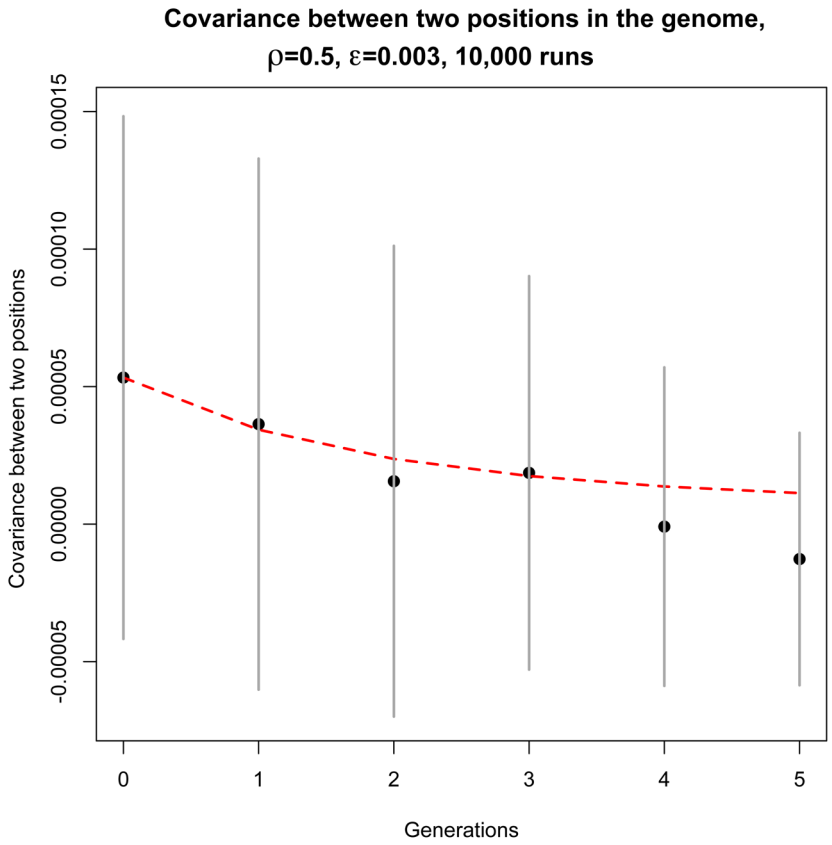

Figure 5.

Comparison between theory (red dashed line) and simulation (black dots) of the contribution of each position pair (denoted x and y, which here we took to be one third and two thirds of the genome) to the variance of HD_0 for an exponentially growing population. Here $\rho = 0.5$ and $\varepsilon = 0.003$. These results were obtained from 10,000 independent simulation runs with the same initial conditions, as described in the main text. The simulation is represented by the black dots, and the theory by the dashed red line. Confidence intervals (±1.96 times the standard deviation over runs) are represented by vertical gray segments.

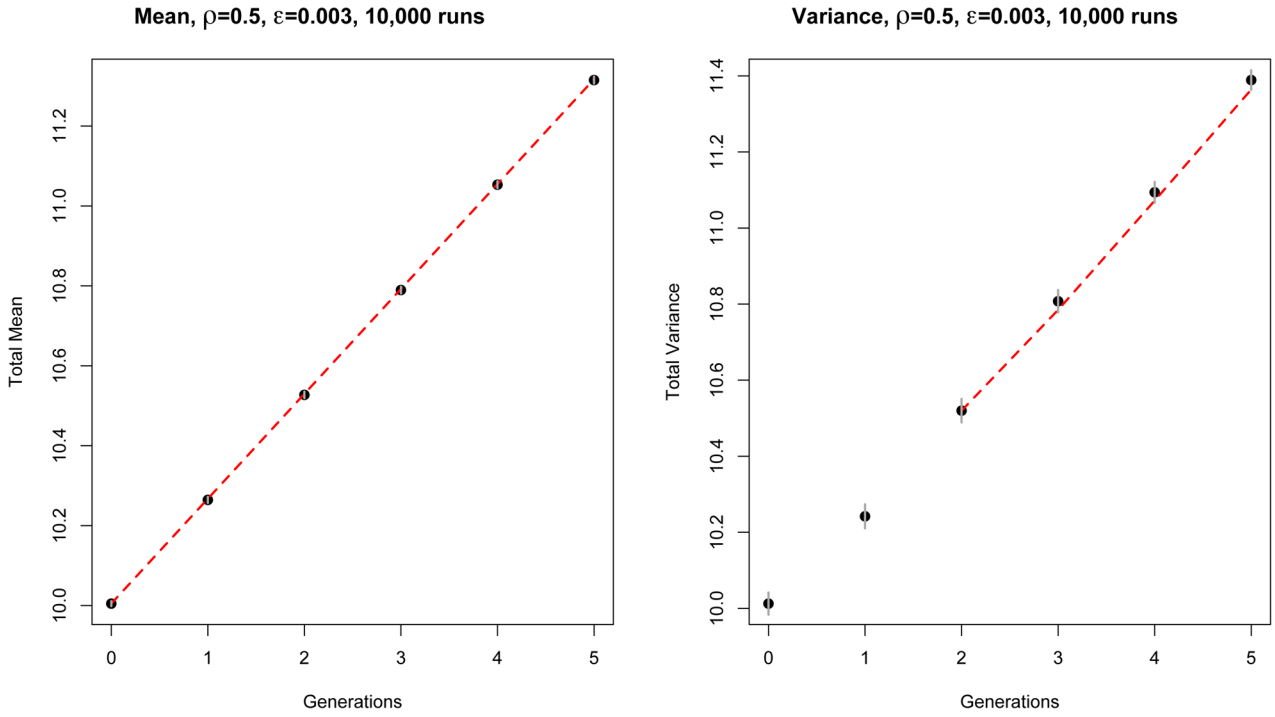

Figure 6.

Comparison between theory (red dashed line) and simulation (black dots) of the mean and variance HD_0 for an exponentially growing population with $\rho = 0.5$ and $\varepsilon = 0.003$. The left panel represents the mean, and the right panel the variance, both averaged over 10,000 runs. Like before, at generation $g = 0$ the initial population is made of 100 sequences with HD mean and variance $\lambda_0 = 10$. The simulation is represented by the black dots, and the theory by the dashed red line. Confidence intervals (±1.96 times the standard deviation over runs) are represented by vertical gray segments. In the right panel, the first few generations are omitted. The theory we developed neglects early stochastic events, which are of the order $\frac{1}{N_g}$, where N_g is the population size at generation g. Hence, we find that only when the population size is large enough, do theory and simulation overlap.

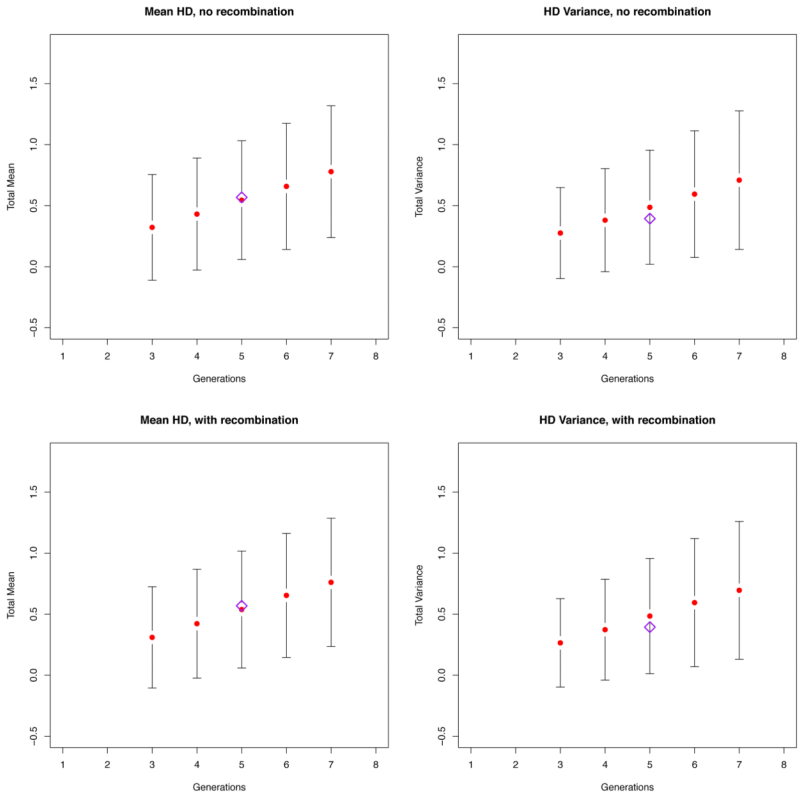

Figure 7.

Simulation runs based on the patient SUMA (Keele et al., 2008). Assuming a single strain initiated the infection, at every generation we sampled $N = 35$ sequences, calculated the HD mean and HD variance, and then repeated 10,000 times. The top panels represent the simulation run in the absence of recombination, the bottom panels represent the simulation when recombination was present ($\rho = 2 \times 10^{-5}$, as estimated in Neher et al. (2010)). The red dots represent the HD mean (left) and HD variance (right) averaged over 10,000 runs. The black vertical bars represent the 95 % CIs. Finally, the purple diamond denotes SUMA's HD mean (left) and HD variance (right). As one can see, the effect of recombination is so small that both models fit the data well, thus validating our previous approach.