Journal of Theoretical Biology 261 (2009) 341–360

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Review

Modeling sequence evolution in acute HIV-1 infection

Ha Youn Lee [a,b,1], Elena E. Giorgi [a,c,1], Brandon F. Keele [d], Brian Gaschen [a], Gayathri S. Athreya [a], Jesus F. Salazar-Gonzalez [d], Kimmy T. Pham [d], Paul A. Goepfert [d], J. Michael Kilby [d,2], Michael S. Saag [d], Eric L. Delwart [e], Michael P. Busch [e], Beatrice H. Hahn [d], George M. Shaw [d], Bette T. Korber [a,f], Tanmoy Bhattacharya [a,f], Alan S. Perelson [a,*]

[a] *Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM 87545, USA*
[b] *Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY 14642, USA*
[c] *University of Massachusetts, Amherst, MA 01002, USA*
[d] *University of Alabama at Birmingham, Birmingham, AL 35223, USA*
[e] *Blood Systems Research Institute, San Francisco, CA 94118, USA*
[f] *Santa Fe Institute, Santa Fe, NM 87501, USA*

ARTICLE INFO

Article history:
Received 16 April 2009
Received in revised form
20 July 2009
Accepted 29 July 2009
Available online 4 August 2009

Keywords:
HIV-1
Population dynamics
Viral evolution

ABSTRACT

We describe a mathematical model and Monte Carlo (MC) simulation of viral evolution during acute infection. We consider both synchronous and asynchronous processes of viral infection of new target cells. The model enables an assessment of the expected sequence diversity in new HIV-1 infections originating from a single transmitted viral strain, estimation of the most recent common ancestor (MRCA) of the transmitted viral lineage, and estimation of the time to coalesce back to the MRCA. We also calculate the probability of the MRCA being the transmitted virus or an evolved variant. Excluding insertions and deletions, we assume HIV-1 evolves by base substitution without selection pressure during the earliest phase of HIV-1 infection prior to the immune response. Unlike phylogenetic methods that follow a lineage backwards to coalescence, we compare the observed data to a model of the diversification of a viral population forward in time. To illustrate the application of these methods, we provide detailed comparisons of the model and simulations results to 306 envelope sequences obtained from eight newly infected subjects at a single time point. The data from $\frac{6}{8}$ patients were in good agreement with model predictions, and hence compatible with a single-strain infection evolving under no selection pressure. The diversity of the samples from the other two patients was too great to be explained by the model, suggesting multiple HIV-1-strains were transmitted. The model can also be applied to longitudinal patient data to estimate within-host viral evolutionary parameters.

Published by Elsevier Ltd.

Contents

* Corresponding author. Tel.: +1 505 667 6829; fax: +1 505 665 3493.
E-mail address: asp@lanl.gov (A.S. Perelson).
[1] Contributed equally.
[2] Current affiliation: Medical University of South Carolina, Charleston, SC, USA.

0022-5193/$ - see front matter Published by Elsevier Ltd.
doi:10.1016/j.jtbi.2009.07.038

1. Introduction

The HIV-1 population in a chronically infected individual is subject to continuous immune selection (Richman et al., 2003; Wei et al., 2003), and evolves to become a complex set of related viruses, often referred to as a quasispecies, through the course of an infection (Lee et al., 2008; Shankarappa et al., 1999; Wolinsky et al., 1996). A reduction in viral complexity at transmission was originally noted in the context of mother to infant transmission (Wolinsky et al., 1992), and has been extensively studied in recent years (Derdeyn et al., 2004; Dickover et al., 2006; Edwards et al., 2006; Painter et al., 2003). During sexual transmission of HIV-1, a genetic bottleneck usually occurs since a limited number of viral strains are transmitted from the complex quasispecies typically found in a donor (Delwart et al., 2001; Derdeyn et al., 2004; Zhang et al., 1993), although other studies have found multiple transmitted variants at a relatively high frequency (Long et al., 2002; Ritola et al., 2004; Sagar et al., 2004; Vernazza et al., 1999). Infection with multiple genetic variants has been associated with genital tract ulcers and use of hormonal contraceptives (Sagar et al., 2004). Recent studies have identified patients in the earliest weeks of infection, many prior to the selective pressure imposed by the newly infected host's initial immune response (Abrahams et al., 2009; Keele et al., 2008; Salazar-Gonzalez et al., 2008). Sequence data from the HIV-1 *env* gene collected during acute HIV-1 infection in 102 subjects by Keele et al. (2008) show a wide range of diversity, with the average number of bases differing between pairs of sequences from the same patient varying between 0.01% and 2.18%, suggesting that both single and multiple viral strain transmissions may have occurred in this cohort.

In this paper, we develop simple models of HIV-1 evolution early in infection with the aim of quantitatively assessing whether infections were established by single or multiple viral strains. Further, in the case of single strain, i.e., homogeneous, infections we aim to identify the most-likely initiating strain or a close descendent that gave rise to the observed lineage.

We derive analytical results and approximate formulas. We used Monte Carlo (MC) simulations to capture the randomness in early HIV-1 evolution and to compare with our analytical results. The analytical results were derived from idealized models whereas the MC simulations allowed for more accurate models that incorporate unequal base composition and an evolutionary based substitution matrix that defines the frequency at which base i is replaced by base j during a mutation event. Previously, Monte Carlo methods have been used to study the within-host dynamics of HIV-1 infection (Heffernan and Wahl, 2005; Kamina et al., 2001; Ribeiro and Bonhoeffer, 1999; Ruskin et al., 2002; Tan and Wu, 1998; Tuckwell and Le Corfec, 1998), but here our focus is on sequence evolution and not viral and T cell dynamics as in these earlier works. Keele and colleagues (2008) applied a variant of this model to the analysis of 102 B-clade infected patients and of 18 experimentally SIV-infected rhesus macaques (Keele et al., 2009). Abrahams et al. (2009) have used the same techniques to analyze a set of 69 C-clade infected subjects.

In this study we provide a complete mathematical description of the model, and explore the implications of varying the baseline assumptions, the input parameters, as well as purely analytical versus computational outcomes. We use just eight of the patients described in Keele et al. (2008) to illustrate nuances in the application of our model. These eight subjects were chosen to be representative of the full set of 102 patients with 80% being characterized as having homogenous infection and 20% heterogeneous infections.

The main problem we focus on here is developing a systematic, reasoned way to determine whether a single strain or multiple stains of HIV-1 infected an individual. This is not straight-forward; even if an individual is infected by a single strain, with time this transmitted virus will diversify. Thus, given a set of sequences one needs to compute how much diversity would be expected by a given time from infection. For sequences collected early after infection, if the diversity is much greater than what is expected from a homogeneous infection, then multiple strain infection is a likely explanation. Other signatures of multiple strain infection might also be present, e.g., the sequences cluster into groups with very different most recent common ancestors. Our model also provides an estimate of the time from the origin of the most recent common ancestor (MRCA) of the sampled variants. If the sampled variants are representative of a lineage that was initiated by the infecting strain, the time to the MRCA should correspond to the time of infection; if the lineage arose in the donor, the time to the MRCA would be longer than the infection; if the lineage arose post-infection in the newly infected individual as a consequence of selection, the time estimated to the MRCA would be less than the time from infection.

To characterize HIV-1 evolution in samples in the earliest weeks of infection, we have modeled the events that occur between virus transmission and peak replication, approximately 21–35 days later. Our model assumes random drift prior to the initial immune response and exponential viral expansion prior to peak viremia. We ignored the effects of selection-based on the premise that the sequences we analyze were obtained sufficiently early in infection that immune responses would not yet provide substantial selective pressure. Although other sources of selection may be present, these also are ignored in our model. The effects of recombination were not modeled either. Comparing the expectations based on the mathematical model and the simulations to

Table 1
Fiebig stage classification for sub-stages of HIV-1 primary infection, and the average and cumulative duration of each phase.

Stage	Duration of each phase (days)	Cumulative duration (days)
Eclipse	10 (7,21)	10 (7,21)
I (vRNA+)	7 (5,10)	17 (13,28)
II (p24Ag+)	5 (4,8)	22 (18,34)
III (ELISA+)	3 (2,5)	25 (22,37)
IV (Western Blot ±)	6 (4,8)	31 (27,43)
V (Western Blot +, p31−)	70 (40,122)	101 (71,154)
VI (Western Blot +, p31+)	Open-ended	

The duration of each stage is presented in Keele et al. (2008). We obtained 95% CIs, shown in parenthesis, by recalculating the confidence intervals presented by Fiebig et al. (2003), through a quadrature method in the following manner. We assume that errors in estimating the length of each stage are uncorrelated, and that they add in quadrature in the cumulative, as defined in Taylor (1982). We use 7–21 days as the duration of the eclipse phase (Keele et al., 2008), and add 2 days with no further error to the duration of the first Fiebig stage because of the increased sensitivity of current assays over the ones available when Fiebig et al. published their work.

observed acute infection sequence data enables us to explore the validity of these assumptions in real infections. We show that under the hypothesis of exponential growth in the infected cell population the frequency distribution of the genetic distances between pairs of HIV-1 sequences follow an approximate Poisson distribution and star-phylogeny topology. This was first shown by Slatkin and Hudson (1991) in the context of the evolution of mitochondrial DNA.

Since the precise time of initial HIV-1 infection cannot usually be known with certainty, the status of acute/early subjects can be classified using the "Fiebig" staging system (Table 1), which is based on an orderly appearance of viral RNA, antigen and antibodies in plasma during early infection (Fiebig et al., 2005, 2003). Prior to stage I, where plasma viral RNA first becomes detectable, is the eclipse phase. The length of this period can be roughly estimated based on clinical histories from a series of studies (Clark et al., 1991; Gaines et al., 1988; Lindback et al., 2000a, 2000b; Schacker et al., 1996) and suggests an average length of about 10 days (range 7–21 days, see Table 1). For each patient in our dataset the Fiebig stage is known. This provides a rough estimate of the time since infection, which can be compared with the estimated time since the MRCA computed with our models from the sampled sequences.

Coalescent theory can also be used to find common ancestors and estimate the time to coalescence. Further, results have been obtained using coalescent theory for the type of linear birth death process that underlies the model described in this paper (Kuhner et al., 1998; Rannala, 1997). When the infection is indeed homogeneous and it meets our model assumptions (exponential viral growth, no selection, etc.), we are under a very particular evolutionary scenario that allows us to use more direct and computationally efficient methods to derive the approximate time to the MRCA. Even though appropriate, Bayesian estimation methods (Drummond et al., 2005, 2006; Kuhner and Smith, 2007), as the ones provided by the software BEAST by Drummond and Rambaut (2007), are computationally intensive even when the number of sequences is small as they perform a whole suite of additional tests that are redundant for our particular evolutionary scenario. The methods presented below are computationally efficient, and our model results can be obtained in minutes for the entire dataset, while in our hands BEAST runs for a single patient took over 3 h. Further, as we shall show, the estimated time to the MRCA using BEAST and using our method are very similar. Thus, we believe the methods we present below enable a rapid exploration of the implications of a forward evolution model, provide consistent results with a coalescence model, and allow us to compare the model to the data to infer biologically interesting information from the sample.

2. Results

2.1. Mathematical models and Monte Carlo simulations

HIV-1 can be transmitted from one individual to another either through the transmission of virus particles or infected cells. Since transmission of virus will quickly generate infected cells, we describe the transmission as if it occurred through infected cells.

In the case of sexually transmitted HIV-1, most often only a single viral sequence initiates the new infection, or only a single sequence grows out to yield a detectable level of viremia (Delwart et al., 2001; Derdeyen et al., 2004; Zhang et al., 1993; Abrahams et al., 2009).Thus, we shall model the case in which infection is initiated by a single HIV-1 sequence. The simplest implementation of this hypothesis is to assume that a single cell, carrying this sequence as a provirus, starts the infection. We shall also examine the case in which multiple cells carrying identical sequences start the infection. When the predictions of these models fail to explain the extent of sequence heterogeneity observed, this is evidence that either multiple sequences have been transmitted, or that immune or other selective pressures are driving the observed level of diversification (Long et al., 2002; Ritola et al., 2004; Sagar et al., 2004; Vernazza et al., 1999).

We assume that the initial infected cell produces virus; most are cleared, but some successfully infect a new generation of cells. The number of secondary infections caused by one infected cell placed in the population of cells of an uninfected individual and fully susceptible to infection is called the basic reproductive ratio, R_0. While data on early viral kinetics in humans is limited, the available data in humans infected with HIV-1 and in monkeys infected with SIV and SHIV show that the virus grows exponentially until a viral load peak is attained a few weeks after infection (Little et al., 1999; Nowak et al., 1997; Stafford et al., 2000). Following the peak, viral levels decline and establish a set-point. At the set-point each infected cell, on average, successfully infects one other cell during its lifetime. If it infected more than one other cell viral levels would increase, whereas if it infected less than one cell viral levels would fall. Thus, once the set-point is established we can assume that the number of infected cells remains constant.

In our model we assume a homogeneous infection in which the virus grows exponentially with no selection pressure, no recombination, no occurrence of back mutations and a constant mutation rate across positions and across lineages. We also assume the virus grows with a fixed generation length and that each infected cell produces the same number of progeny. Clearly, the approximation of exponential growth must break down as target cells become limited and the number of cells infected per generation must decrease from R_0, stabilizing at 1 when the viral set point is established. However, we present evidence indicating that what is important in determining viral diversity is the number of reverse transcription events that have occurred along a lineage, i.e., the number of generations that separate the initial infection from the time at which sequences were obtained, and not the number of cells infected at each generation. In our first, simplest model, we also assume that the infection is synchronous, and thus can be characterized by discrete generations, with the virus infecting exactly R_0 new cells at each generation.

Let all genomes be of same length N_B, let ε be the reverse transcriptase point mutation rate, which we assume to be constant throughout the genome, and let s_0 denote the infecting strain. Then, after the first replication cycle, the R_0 daughter cells will each differ from the infecting strain at exactly n bases (positions) with a probability given by

$$P(mutations = n) = \binom{N_B}{n}\varepsilon^n(1-\varepsilon)^{N_B-n} \equiv Binom(n; N_B, \varepsilon). \quad (1)$$

After the second replication cycle, the total number of mutations in the new generation of infected cells will follow the probability distribution given by

$$P(mutations = n|gen. = 2) = \sum_{k=0}^{n} Binom(n-k; N_B, \varepsilon)Binom(k; N_B, \varepsilon) = Binom(n; 2N_B, \varepsilon).$$

In other words, n mutations in the second generation are the result of all possible combinations of k mutations occurring in the first replication, and $n-k$ occurring in the second, for all integers k between 0 and n. Mathematically, this is given by the convolution of two binomial distributions with the same probability parameter (ε, in our case), which is in turn a binomial distribution (Casella and Berger, 1990). One can see that after a replication cycles,

$$P(mutations = n|gen. = a) = Binom(n; aN_B, \varepsilon).$$

Let HD_0 denote the Hamming distance, i.e., the number of base differences from the infecting strain s_0. In general, given n mutations, we have $HD_0 \le n$, as some of the n mutations may occur at the same site. However, after a replication cycles, the probability that one site has mutated at most once is

$$P = (1-\varepsilon)^a + a\varepsilon(1-\varepsilon)^{a-1}. \quad (2)$$

Therefore, the probability that at least one site across the entire genome has mutated more than once is $Q = 1 - P^{N_B}$, which is approximately $\frac{1}{2}N_B a(a-1)\varepsilon^2$ when ε is small.

For the HIV-1 base substitution rate ε, we use the value of 2.16×10^{-5} base substitutions per replication cycle. The base substitution rate we use is derived from the results of Mansky and Temin (1995), where after a single round of HIV-1 replication they found an overall mutation rate of 3.4×10^{-5}, but this included insertions and deletions. Restricting their results to only base substitutions, we calculated from their data a base substitution rate of 2.16×10^{-5} per base per replication. The sequences that we analyze later are envelope gene sequences with $N_B\sim2600$ bases. Substituting these values into Q, we find that the probability of two or more mutations at the same site increases with the number of replication cycles, yet even for $a = 50$ one finds $Q < 0.0015$. Hence we expect that we can ignore back mutations when we analyze patient sequence data obtained early in infection.

Under this assumption, HD_0 coincides with the number of mutations and hence it follows the same binomial distribution:

$$P(HD_0 = d|a) = Binom(d; aN_B, \varepsilon) + O(a^2\varepsilon^2 N_B) \approx \binom{aN_B}{d}\varepsilon^d(1-\varepsilon)^{aN_B-d}. \quad (3)$$

Notice that this formula assigns a non-zero probability to the Hamming distance being larger than the total number of bases N_B, and this is an artifact of our approximations. In fact, using Chebyshev's inequality (Feller, 1957), one can show (see Materials and methods) that $P(mutations > N_B|a) \le O(a^2\varepsilon^2/N_B) \ll 1$, and thus indeed, within our approximation, $P(mutations > N_B) \approx 0$.

2.2. Synchronous infection, mathematical model

In the synchronous model, we assume that each cell infects R_0 other cells at the end of its life cycle. At generation a from the infection event, there will be R_0^a infected cells, each one carrying one HIV-1 genome. Further, at generation a

$$P(HD_0 = d, a) = \binom{aN_B}{d}\varepsilon^d(1-\varepsilon)^{aN_B-d}. \quad (4)$$

In order to analyze single time point data, it is useful to compute the intersequence HD distribution rather than HD_0 (since for the latter we would need to assume that we know the infecting sequence s_0). In other words, rather than comparing any given sequence to the founder sequence, we take all possible sequence pairs in the sample and compute their relative HD. Given two sequences, s_1 and s_2, picked at random at time t, that have evolved independently from the common ancestor s_0, the distribution of HD[s_1,s_2] is

$$P(HD[s_1,s_2] = d|a) = \sum_{k=0}^{d} P(HD_0 = k|a)P(HD_0 = d-k|a) = \sum_{k=0}^{d} Binom(k; aN_B, \varepsilon)Binom(d-k; aN_B, \varepsilon) = Binom(d; 2aN_B, \varepsilon). \quad (5)$$

The assumption of independent evolution of sequences s_1 and s_2 from the common ancestor s_0 is equivalent to assuming that the phylogeny of these sequences is star-like with s_0 being the only common ancestor. However, it is also conceivable that the common ancestor of s_1 and s_2 is not the transmitted sequence s_0 but rather a sequence that evolved from s_0. This is an important issue, as the identity of the transmitted sequence is generally not known. As illustrated in Fig. 1, the probability of s_1 and s_2 coalescing m generations from the ancestor, with $m < a$, is the number of all sequences coalescing m generations from the ancestor (minus the one we have already picked), divided by the total number of sequences at generation a (again, minus the

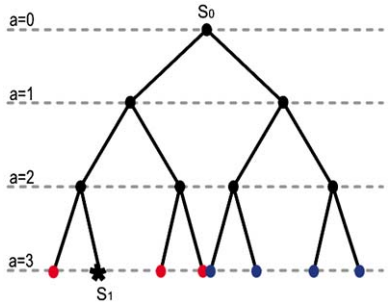

Fig. 1. Probability of early coalescence when $R_0 = 2$. Let s_1 be the randomly picked sequence at generation $a = 3$. The common ancestor between s_1 and any of the sequences represented by the blue dots is indeed s_0 (the founder strain), whereas the common ancestor between s_1 and any of the sequences represented by the red dots sits at generation $a = 1$. There are 2^2-1 "red" sequences out of a total of 2^3-1 sequences we could choose s_2 from at generation 3. Therefore, the probability of s_1 and s_2 coalescing at generation $m = 1$ instead of generation 0 is $2^2 - 1/2^3 - 1 = 3/7 = 0.43$. This probability decreases exponentially as m gets larger. Indeed, by generalizing the above argument, one can see that if s_1 is picked at random at generation a, there are exactly $R_0^{a-m} - 1$ sequences that coalesce m generations after the founder strain, out of the total of $R_0^a - 1$ sequences s_2 could be picked out of. Hence the probability of coalescing $a-m$ generations back, with $1 \le m < a$, is given by $P = R_0^{a-m} - 1/R_0^a - 1 \approx O(R_0^{-m})$, and the probability of coalescing at the founder strain is $R_0^{a-1}/R_0^a - 1 = O(1/R_0)$. So, the larger R_0, the smaller the error in assuming that everything coalesces at the founder.

one we picked already)

$$P(MRCA[s_1,s_2]\leq a-m|s_1,s_2\in x)=\frac{R_0^{a-m}-1}{R_0^a-1}. \quad (6)$$

As shown in Materials and methods, this probability approaches zero rapidly as a and m get large.

We can use Eq. (6) to determine the probability that two sequences from a given individual coalesce at a time point other than at the transmitted sequence. For example, for samples obtained a few weeks into the infection and under no selective pressure, e.g., $a=5$ and $R_0=6$ (see below) the probability of coalescence later than generation 0, from Eq. (6) is 0.167 for $m=1$, 0.028 for $m=2$ and 0.0045 for $m=3$. For greater values of a, the probability of coalescing at generation 3 or higher is <0.005. Thus, we expect that use of Eq. (5) and the assumption of a star-like phylogeny to hold for all patient samples in which a single viral strain was transmitted. However, even for large a, the probability of coalescing 1 or 2 generations away from the founder strain still remains non-negligible: even at generation 40, the probability of coalescence later than generation 0, is still 0.167 for $m=1$, and 0.028 for $m=2$. In a set of sequences from an individual, these estimates of $<0.5\%$ chance of being three generations away from the actual founder, a 3% chance of being two generations away, and a 17% chance of being 1 generation away hold for each independent pair of comparisons.

Samples that diverge from a star phylogeny often show patterns of shared mutations across sequences. The above calculation shows that coalescence not at the founder strain is most likely one generation away from the actual founder. Therefore, a mutation observed very early in the infection, could possibly carry on and develop as a second lineage. Also, we cannot rule out that the founders of these two lineages were simply transmitted from the donor and hence coalesced in the donor. For infections transmitted during the acute phase before much diversification occurred in the donor this may even be likely. Lastly, we observe that since the base substitution rate ε is so small it is very likely that the sequences at $a=0$ and 1 are the same. In fact, with probability $(1-\varepsilon)^{R_0N_B}$ all R_0 proviruses at generation 1 are identical to the founder; e.g., with $R_0=6$ and $N_B=2600$ this probability is 0.71.

For ease of notation, let the intersequence Hamming distance distribution be denoted HD_I. Thus,

$$P(HD_I=d|a)=P(HD[s_1,s_2]=d|a)=Binom(d;2N_Ba,\varepsilon). \quad (7)$$

We also define some basic quantities based on the calculated Hamming distance distribution from the initial strain and the intersequence Hamming distance distribution between all possible sequence pairs. The divergence at generation a is defined as the average Hamming distance per base from the initial founder strain. From Eq. (4), this can be estimated as the mean of the binomial distribution divided by the number of bases, and is εa. The diversity at generation a is the average intersequence Hamming distance per base between sequence pairs and from Eq. (7), diversity $=2\varepsilon a$. The variance of the intersequence HD distribution at generation a divided by the number of bases is $2\varepsilon a(1-\varepsilon)$. We also introduce the % sequence identity, defined as the proportion of sequences identical to the sequence, s_0, of the infecting strain. As the virus evolves, the population diversifies from the founder strain and the proportion of the total population that is identical to the MRCA decays exponentially. By setting $d=0$ in Eq. (4), we have % sequence identity $=Binom(0;aN_B,\varepsilon)=(1-\varepsilon)^{aN_B}$.

Finally, we wish to compute the expected maximum HD_I. This is motivated by the observation that more extensive diversity is seen in some transmission cases, presumably the result of infections established by two or more distinct HIV-1 strains (Long et al., 2002; Ritola et al., 2004; Sagar et al., 2004; Vernazza et al., 1999). Thus, a strategy for systematic classification of homogeneous, single strain infections versus heterogeneous infections is needed. To address this, we ask what is the maximum diversity that one would expect in a single strain infection within a time frame compatible with the subject's Fiebig stage at the time of sampling. The maximum diversity expected depends on the sample size: the larger the sample, the higher the expected maximum, as outliers become more likely to be sampled. The expected maximum HD at generation a is computed in the Materials and methods.

2.3. Monte Carlo simulation of the synchronous infection model

The mathematical model derived above assumes that all mutations have equal probability. However, in HIV-1 the four bases are not equally frequent and the rates of base substitution are not all equal. Further, we neglected more than one mutation occurring at the same site, although this potentially can occur. To examine the influence of these biological details on the predictions of HIV-1 sequence evolution, and to capture stochastic effects, we developed a simple Monte Carlo simulation and compared its predictions with those of the mathematical model.

The simulation starts at "generation 0" with one cell infected with a single HIV-1 sequence N_B bases long. The initial sequence is generated randomly with base frequencies of 0.19 (T and C), 0.37 (A) and 0.24 (G), based on typical frequencies measured in the full-length envelope gene (see Materials and methods). We assume that the initially infected cell infects R_0 other cells synchronously. We chose $R_0=6$, representing the average value of R_0 across 10 different patients studied by Stafford et al. (2000). During the infection of these cells mutations can occur. In the simulation, the sites for the occurrence of the base substitutions are chosen randomly from the N_B bases. The probability that base i is replaced by j is given by a 4×4 transition matrix deduced from a maximum likelihood general time reversible (GTR) model of substitutions that occur in the full length HIV-1 envelope gene (for the matrix and the details see Materials and methods). Unlike in our analytical model, we allow the same site to mutate more than once. For a given sample size N_S, the MC simulation produces an HD frequency distribution for each generation step within a certain time frame chosen by the user. At each generation exactly N_S sequences are sampled (with replacement) and the intersequence HD distribution is thus constructed.

We implemented the MC code so that the user has the option of running either one or multiple iterations. When multiple iterations are done, the HD frequencies computed at each generation step are averaged over all iterations, thus retrieving the mean field description described by the analytical method. On the other hand, when not averaging over all iterations, one captures the stochastic effects due to rare mutations, e.g., the run-to-run variance, which, in real life, corresponds to the host-to-host variation. This is completely neglected in the analytical method, which is a mean field description.

In the synchronous model the population of infected cells grows exponentially as R_0^a. Once the number of infected cells reaches 10^4, which is close to some estimates of the effective population size N_e of the virus in a mature infection (Achaz et al., 2004; Leigh Brown, 1997; Wakeley, 2008), there is no need to expand the population further in order to study the diversity of genetic sequences as long as we allow the proviral sequences contained in these 10^4 to continue to evolve. We thus cap the total population of infected cells at 10^4 and let those cells produce 10^4 R_0 progeny. We then randomly sample 10^4 cells out of the 10^4 R_0, and continue the simulation. In this way, we are able to maintain

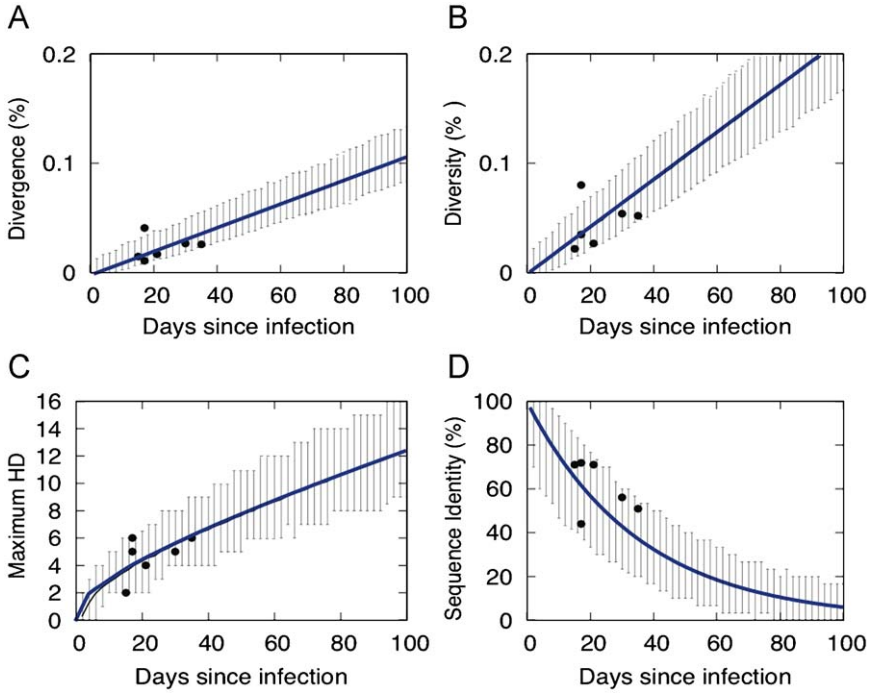

Fig. 2. Synchronous infection model. Dynamics of (A) divergence, (B) diversity, (C) maximum HD_l, and (D) the % sequence identity for sequences of length 2600, derived from mathematical model (blue line) and from 10^3 Monte Carlo simulations. The corresponding values for the six homogeneous patients are also given (filled circles). In each Monte Carlo simulation a sample of $N_S = 30$ randomly drawn sequences was used to generate the plot and the 95% confidence intervals (black lines). The generation time was assumed to be 2 days in order to convert the results to days since infection. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

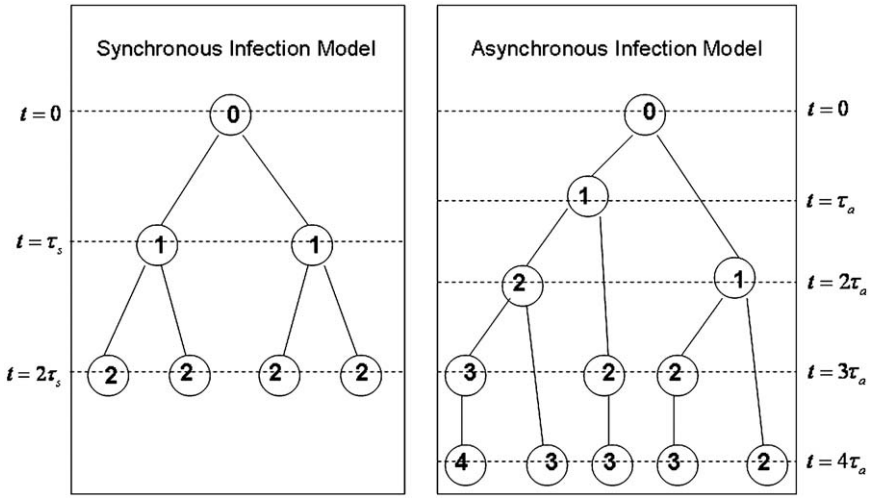

Fig. 3. Schematic diagrams of the synchronous and asynchronous MC infection models. Each circle represents one infected cell, and the number within the circle represents the age of the infected cells' viral genome measured in terms of the number of times it has been reverse transcribed since the founder strain. The generation time in the synchronous model is τ_s and τ_a denotes the time the first infected progeny are produced in the asynchronous model (defined in the text).

the population of infected cells constant over the rest of the simulation, which is equivalent to what occurs when the viral set-point is reached. We confirmed that our simulation results (i.e., means and 95% CIs) did not depend on the cutoff value of infected cells as long as it is above 10^3 (data not shown).

Using the MC simulation, we computed divergence, diversity, maximum HD and % sequence identity (defined in the previous section) as functions of time, and confirmed that their dependence on the number of generation steps matched the dependency found analytically. Fig. 2 shows this comparison as well as 95% confidence intervals computed from 1000 MC simulations using a typical number of sampled sequences, $N_S = 30$.

2.4. Asynchronous infection

The synchronous infection model is a simplification. In reality, cell infections occur at random times by the viruses released from an infected cell. Viral production starts on average about 24 h after a cell is initially infected (Perelson et al., 1996), and most likely continues until cell death. In order to understand the dependence of our results on the synchronous approximation, we devised a second model in which infections are asynchronous. Again we will assume each cell infects R_0 others during its lifetime. While each of the R_0 infections could occur at different times, here we take a first step in assessing the role of asynchrony by assuming the

infections occurs at two times. Each infected cell infects α cells at an intermediate time, and the rest at the end of its life-time. Even though this scenario is again a simplification, it is useful to analyze the differences from the previous model, and as we shall show below it preserves the same basic mathematical structure as before. However, because an infected cell contributes progeny to the next generation before the end of its life there is now a faster growth rate of the infected cell population.

Let τ_a and $2\tau_a$ be the times at which a newly infected cell infects other cells, and suppose it infects α cells at τ_a, and γ at $2\tau_a$, with $\alpha+\gamma = R_0$, the total number of cells one cell infected during its lifetime. Notice that if $\gamma = 0$, then infections only occur at one time, $\alpha = R_0$, and τ_a is the generation time. Hence this model reduces to the synchronous model with generation time $\tau_s = \tau_a$, where τ_s denotes the duration of a single generation step in the synchronous model.

Denote by $I(a,t)$ the number of infected cells of age a at time t. Here, by age, we mean the age of the infecting genome, measured in terms of the number of times the viral genome has undergone reverse transcription, rather than the cell's physical age. Thus, a is equivalent to the generation variable we used in the synchronous model. As a consequence, at any given time t, under the synchronous model all cells belong to the same age group, whereas in the asynchronous model, the infecting genomes have undergone a minimum of $\lceil t/2\tau_a \rceil$ replication cycles, and a maximum of $\lfloor t/\tau_a \rfloor$ (see Fig. 3), where $\lceil\ \rceil$ and $\lfloor\ \rfloor$ are the ceiling and floor functions defined as $\lfloor x \rfloor = \max\{n \in Z | n \le x\}$ and $\lceil x \rceil = \min\{n \in Z | n \ge x\}$, where Z is the set of integers.

Let I_0 denote the initial number of infected cells and let $n = \lfloor t/\tau_a \rfloor$, i.e., n represents time measured in units of τ_a. In the following we shall use n as the time variable. A general expression for $I(a,n)$ is derived in Materials and methods, and is given by

$$I(a,n) = I_0 \alpha^a \left(\frac{\gamma}{\alpha}\right)^{n-a} \binom{a}{n-a} \text{ for } a = \lceil \frac{n}{2} \rceil, \ldots, n \text{ and } 0 \text{ otherwise.} \tag{8}$$

As illustrated in Fig. 3, at any given time n, the HIV-1 strains in an individual have undergone a minimum of $\lceil n/2 \rceil$ and a maximum of n replication cycles. Then, as a function of time, the random variable HD_0 follows the probability distribution:

$$P(HD_0 = d|n) = \frac{\sum_{a=\lceil n/2 \rceil}^{n} \binom{aN_B}{d} \varepsilon^d (1-\varepsilon)^{aN_B - d} I(a,n)}{\sum_{a=\lceil n/2 \rceil}^{n} I(a,n)}. \tag{9}$$

In Materials and methods we show the following relationship:

$$\sum_{a=\lceil n/2 \rceil}^{n} I(a,n) = \frac{I_0}{2\varphi} \left(\frac{\alpha}{2}\right)^n [(1+\varphi)^{n+1} - (1-\varphi)^{n+1}] = I_0 \alpha^n F_n, \tag{10}$$

where $\varphi = \sqrt{1 + 4(\gamma/\alpha^2)}$ and $F_n = (1+\varphi)^{n+1} - (1-\varphi)^{n+1}/2^{n+1}\varphi$. By substituting Eq. (10) into Eq. (9) we obtain

$$P(HD_0 = d|n) = \frac{1}{F_n} \sum_{a=\lceil n/2 \rceil}^{n} \binom{a}{n-a} \binom{aN_B}{d} \left(\frac{\gamma}{\alpha}\right)^{n-a} \varepsilon^d (1-\varepsilon)^{aN_B - d}, \tag{11}$$

with mean $\mu_n = \lambda_n \varepsilon N_B$ and variance $\sigma_n^2 = \lambda_n \varepsilon (1-\varepsilon) N_B$, where

$$\lambda_n = n \frac{1+\varphi}{2\varphi} + \frac{1-\varphi}{\varphi^2}. \tag{12}$$

Notice, as one would expect, that in neither the synchronous nor the asynchronous model, does the probability distribution depend on the initial number of infected cells, I_0.

We show in Materials and methods that up to terms in $O(\varepsilon^2)$ $P(HD_0 = d|n)$ is a Poisson distribution with mean $\mu_n = \lambda_n \varepsilon N_B$, in other words

$$P(HD_0 = d|n) \approx Pois(d, \lambda_n \varepsilon N_B) + O(\varepsilon^2). \tag{13}$$

As a consequence, given that the intersequence HD distribution, HD_I, is to a good approximation the self-convolution of the HD_0 distribution, we get that HD_I also follows a Poisson distribution with parameter $2\lambda_n \varepsilon N_B$. Therefore, analogously to the synchronous model, divergence $= \lambda_n \varepsilon N_B$, diversity $= 2\lambda_n \varepsilon N_B =$ variance and % sequence identity $= Exp(-2\lambda_n \varepsilon N_B)$. The expected maximum HD is calculated as in the synchronous infection model, Eqs. (8) and (9), except we now insert $P_k^{(a)} = P(HD_0 = k|a)$ from Eq. (11).

In the synchronous model, the intersequence Hamming distance, Eq. (5), follows a binomial distribution, which is also well approximated by a Poisson distribution with equal mean since $2aN_B$ is large, ε is small and $\lambda = 2aN_B\varepsilon$ is of reasonable size (Casella and Berger, 1990). Hence in both scenarios the HD follows essentially the same probability distribution, but with different divergence and diversity, implying a different rate of evolution due to the asynchrony in infection.

Let $N(n)$ be the total number of newly infected cells at time n. For $I_0 = 1$ and $R_0 \gg 1$, we show in Materials and methods that

$$N(n) \approx \left(\frac{\alpha}{2}(1+\varphi)\right)^n = \left(\frac{\alpha}{2}(1+\varphi)\right)^{\lfloor t/\tau_a \rfloor},$$

whereas for the same time t, in the synchronous model, $N(n) \approx (R_0)^{\lfloor t/\tau_s \rfloor}$, where τ_s denotes the generation time in the synchronous model. Notice that $\tau_a \neq \tau_s$ (see Materials and methods), hence the two scenarios both provide descriptions of an exponentially growing infected cell population but with different exponents.

We also devised a partially randomized model, in which the α, instead of being a fixed parameter, is a random number uniformly distributed with mean $R_0/2$, and then γ is always chosen to be $R_0 - \alpha$. This introduces an additional variability the effect of which is inversely proportional to the total population size. Therefore, large fluctuations from the above description will occur within the first 2–3 generation steps, but will be negligible after that, thus recovering the above mean field description.

An important difference between the asynchronous and synchronous models is in the dependence on R_0. In both models,

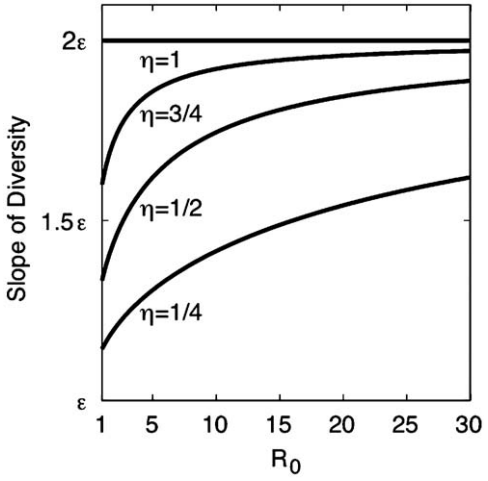

Fig. 4. The slope of the diversity versus time for different values of η and R_0. The diversity depends linearly on the number of generation steps, i.e., diversity $= \varepsilon f(\eta,R_0)n$, where n is the number of generation steps and $0 < \eta \le 1$. As R_0 increases, the slope of the diversity versus time plot increases, and f approaches 2 as R_0 increases.

divergence and diversity depend linearly on the number of generation steps. In the synchronous model, the slope of the diversity is 2ε and it does not depend on R_0. In the asynchronous model, the slope has the additional factor

$$\frac{\lambda_n}{n} = \frac{1+\varphi}{2\varphi}, \tag{14}$$

with $\varphi = \sqrt{1+4(\gamma/\alpha^2)}$ and $\gamma+\alpha = R_0$. One sees that in the asynchronous model there is a dependence on R_0. Let $\alpha = \eta R_0$ for some $0<\eta\leq 1$, and $\gamma = R_0(1-\eta)$. Then, one gets the expression of the slope in diversity as

$$\varepsilon\left(1+\frac{1}{\sqrt{1+4\dfrac{1-\eta}{R_0\eta^2}}}\right). \tag{15}$$

Notice that the when $\eta = 1$, we retrieve as a special case the synchronous model, and indeed Eq. (15) yields exactly 2ε. However, when $\eta \neq 1$, the above expression is strictly smaller than 2ε, and approaches 2ε as R_0 increases. Fig. 4 illustrates the dependence of the slope of the diversity as a function of the number of generation steps on R_0.

Table 2
Analytical formulas for the basic quantities in the synchronous and asynchronous infection models.

	Synchronous infection model	Asynchronous infection model
Divergence	εa	$\varepsilon\lambda_n$
Diversity	$2\varepsilon a$	$2\varepsilon\lambda_n$
Variance	$2\varepsilon a(1-\varepsilon)\simeq 2\varepsilon a$	$2\varepsilon(1-\varepsilon)\lambda_n \simeq 2\varepsilon\lambda_n$
Maximum	$\sum_{M=0}^{N_B} M \times P_{syn}(HD_{\max}=M\|a)$	$\sum_{M=0}^{N_B} M \times P_{asyn}(HD_{\max}=M\|n)$
% Sequence identity	$(1-\varepsilon)^{aN_B}$	$e^{-\lambda_n \varepsilon N_B}$

The expression for λ_n is given by Eq. (12).

Table 2 summarizes the above formulas for both the synchronous and the asynchronous models. In both models, divergence, diversity, and variance increase linearly as a function of time. Diversity and variance are expressed by the same mathematical formula, which is a consequence of the fact that the intersequence HDs follow a Poisson distribution. Divergence, diversity, and HD_I variance do not depend on the number of bases per sequence. In contrast, the maximum HD and the % sequence identity depend on the sequence length N_B. In the synchronous infection model, all the quantities are independent of the basic reproductive ratio, while in the asynchronous model, they do depend on the R_0, as discussed above. Quantitatively, however, the asynchronous infection model provides minor corrections compared to the synchronous infection model.

2.5. Monte Carlo simulation of the asynchronous infection model

We keep the basic framework used for the synchronous MC simulation, and choose parameters $R_0 = 6$, $\varepsilon = 2.16\times10^{-5}$, and $\tau_a = 1.5$ days (see Materials and methods). Furthermore, we simulate the situation where $I_0 = 1$ and each infected cells infects $R_0/2$ cells at τ_a, and the remaining $R_0/2$ at time $2\tau_a$. We also devised a second scenario under which each infected cells infects α cells at time τ_a, where α is drawn from a random uniform distribution with mean $R_0/2$, and $R_0-\alpha$ at time $2\tau_a$. We confirmed that both types of simulations, after generation step 4, are in excellent agreement, and also confirmed our analytical results described in the previous section. Furthermore, both % divergence and % diversity grow linearly with time (in units of τ_a) at rates of 1.79×10^{-5} and 3.57×10^{-5}, respectively (i.e., 1.2×10^{-5} and 2.4×10^{-5} per day, respectively).

At times τ_a, $2\tau_a$, $3\tau_a$, and so on, we sample N_S sequences, construct the HD distribution and compute the divergence, diversity, HD_I variance, % sequence identity and the expected

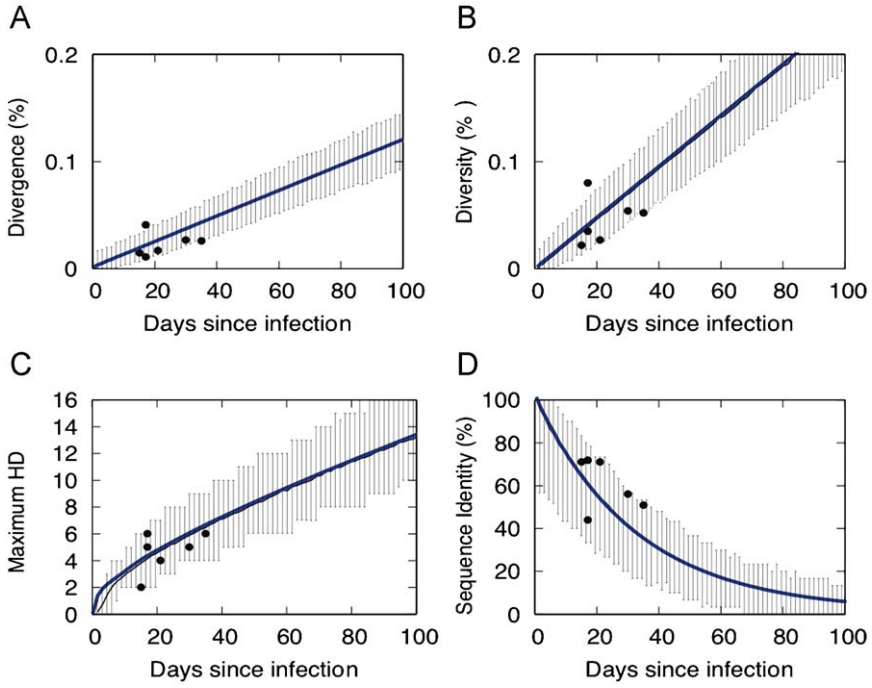

Fig. 5. Asynchronous infection model. Dynamics of (A) divergence, (B) diversity, (C) maximum HD_I, and (D) the % sequence identity for sequences of length 2600, derived from mathematical model (blue line) and from 10^3 Monte Carlo simulations. The corresponding values for the six homogeneous patients are also given (filled circles). In each Monte Carlo simulation a sample of $N_S = 30$ randomly drawn sequences was used to generate the plot and the 95% confidence intervals (black lines). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

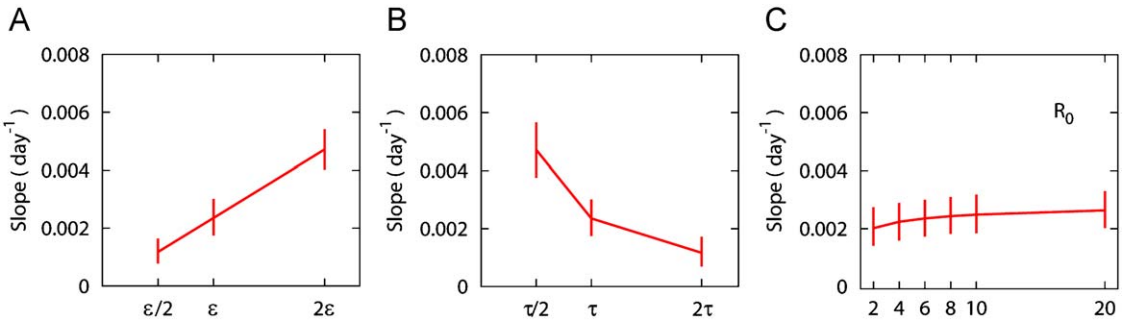

Fig. 6. Parameter dependence (asynchronous infection model). The dependence of the slope of the mean diversity, computed from a sample of $N_S = 30$ sequences at each time, on (A) the base substitution rate, ε, (B) the generation time, τ, and (C) the basic reproductive number R_0. The vertical bars indicate 95% confidence intervals.

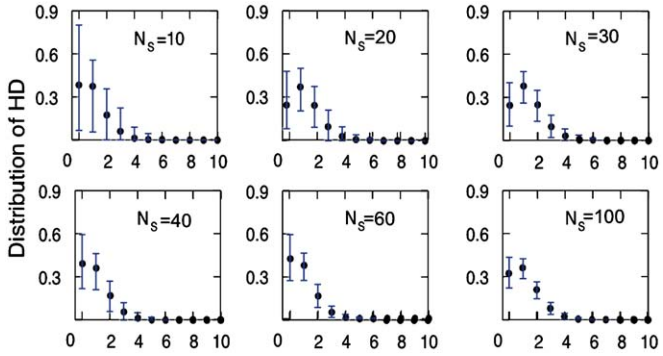

Fig. 7. Sample size dependence of the Poisson fit. Pairwise HD distribution from the asynchronous MC simulation with different numbers of sampled sequences, N_s, from 10 to 100 at day 20. The average frequency of each HD (black dot) and the 95% confidence interval from 10^3 MC runs (blue lines) are plotted as a function of HD. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

maximum HD. As for the synchronous model, we find that the MC results agree well with the analytical results given in Table 2 (Fig. 5).

We examined the dependency of our results on the values chosen for the parameters τ, R_0, I_0 and ε (Fig. 6). In agreement with our calculations (Table 2), the diversity increases linearly with time (not shown) at a rate proportional to ε, inversely proportional to τ_a (Figs. 6A and B), dependent on R_0 (Fig. 6C) and independent of I_0 (not shown). The mutation rate controls the rate of increase of the divergence and the diversity. The larger the mutation rate, the faster the genomes mutate, hence the steeper the growth of the diversity (Fig. 6A). The greater the generation time, the slower the genomes diversify, hence the smaller the growth in diversity (Fig. 6B). The computed diversity depends on R_0 (Fig. 6C). At low values there is a strong dependence, e.g., from $R_0 = 2$ to 6 there is 15.9% increase in the slope of the diversity. On the other hand, for larger values of R_0, the dependence is much weaker (Fig. 6C).

2.6. Sample size dependence

When patients are sampled, only a limited number of sequences are obtained. The Monte Carlo simulations provide a means of assessing the variability introduced by limited sequencing and can be used to choose the number of samples to be sequenced based on the trade-offs between cost and accuracy. Fig. 7 shows the variation in the HD frequency distribution as the sample size N_S varies. Clearly, the larger the sample size, the smaller the variation.

We also asked what is the probability of erroneously classifying an infection as homogeneous due to poor sampling. In other words, what is the chance of missing a second lineage that represents a small percentage of the entire population when N_S sequences are sampled. We assumed a range of hypothetical low population prevalence, from 0.1% to 20%, and computed the probability that, out of a sample size N_S, all sequences come from the dominant lineage. This is given by a binomial probability of getting zero "successes" out of N_S trials, where a "success" is defined as a sequence belonging to the minor lineage, and the probability p of success is given by its population prevalence

$$P = Binom(0; N_S, p). \tag{16}$$

Thus, for example, for $N_S \geq 30$, we can be 95% confident that any missed variant would comprise $<10\%$ of the total viral population (see Fig. pS9 in Keele et al., 2008). The number of sequences sampled from our eight patients ranged from 16 to 50. When $N_S = 16$, we are 95% certain that we have sampled all lineages that represent at least 20% of the population. On the other hand, when $N_S = 50$, we are 95% confident that any missed lineage would have to represent 5% or less of the total population.

2.7. Analysis of sequence data from eight acute patients

Plasma samples were obtained from eight subjects with acute or very recent HIV-1 subtype B infection. Laboratory testing showed that seven subjects were in Fiebig stage II, both viral RNA and p24 antigen were detected, and one subject was in Fiebig stage III, HIV-1 IgM in addition to viral RNA and p24 was detected (see Table 1). Detailed information of each subject is summarized in Table 3. Furthermore, samples were analyzed using the *Hypermut* tool (LANL) to test for overall enrichment of G-to-A hypermutation patterns with APOBEC3G/F signatures (Bourara et al., 2007; Harris and Liddament, 2004; Simon et al., 2005). Because these mutational patterns occur at a faster rate than ordinary mutations, they violate one of our model assumptions (constant mutation rate across sites and lineages). Only one sample, SUMA, presented mutations (three) that were found to carry the APOBEC3G/F signature. When all mutations were kept, the sample did not fit the Poisson model (χ^2 goodness of fit p-value $= 0.04$, where a $p < 0.05$ represents significant departure from the Poisson distribution), whereas a good fit was restored upon removing the APOBEC signatures (χ^2 goodness of fit p-value $= 0.243$) (Keele et al., 2008).

For each of the eight samples we computed the HD_0 and HD_I frequencies, as well as the divergence, diversity, HD_I variance, % sequence identity and maximum HD_I. All quantities are summarized in Table 4. In order to compute the HD_0 distribution, we used the LANL tool *Consensus Maker* to compute a consensus sequence

Table 3
Patient characteristics at time of sampling, number of envelope sequences obtained and their length, and GenBank accession numbers.

Subject	Gender	Fiebig stage	Number of sequences[a]	Base length	Viral load (copies/ml)	CD4 count (cells/ml)	GenBank accession number
WEAU0575	M	II	43	2581	216,415	714	EU577344–EU577387
WITO4160	M	II	16	2550	325,064	247	EU577388–EU577403
TRJO4551	M	II	16	2574	8,121,951	336	EU577101–EU577118
SUMA0874	M	II	35	2568	939,260	760	EU577039–EU577073
1054	M	II	36	2562	320,000	NA	EU575244–EU575282
1056	M	II	46	2589	140,000	NA	EU575283–EU575328
1051	F	III	50	2619	280,000	NA	EU575134–EU575183
BORI0637	M	II	29	2607	2,400,000	902	EU576274–EU576302

[a] Number of sequences used in our analysis. Patients WEAU, TRJO and 1054 had 1, 2 and 3 additional sequences, respectively, that were removed from the analyses because of large internal deletions.

Table 4
Analysis of 8 HIV-1 acute sequence samples using the asynchronous infection model.

Subject	Divergence (%)	Diversity (%)	Variance (%)	Max. HD	Sequence identity (%)	λ (95% CIs)	Estimated time (days) to MRCA from Poisson fit (95% CIs)	Estimated time (days) to MRCA from MC (95% CIs)
WEAU0575	0.026	0.052	0.055	6	51	1.34 (1.28,1.40)	22 (17, 29)	25 (14,36)
WITO4160	0.027	0.054	0.062	5	56	1.38 (1.32, 1.43)	23 (12, 36)	26 (12,44)
TRJO4551	0.041	0.080	0.083	6	44	2.06 (1.99, 2.14)	34 (22, 49)	37 (21,54)
SUMA0874	0.011	0.022	0.015	2	71	0.57 (0.54, 0.65)	10 (6, 13)	14 (3,21)
1054	0.017	0.035	0.048	5	72	0.89 (0.84, 0.94)	15 (9, 22)	16 (6,26)
1056	0.013	0.027	0.027	4	71	0.70 (0.66, 0.74)	11 (7, 16)	13 (5,21)
1051	0.42	0.73	22.4	71	42	NA	NA	NA
BORI0637	1.2	1.7	24.9	77	19	NA	NA	NA

For each subject, we calculated divergence, diversity and maximum intersequence HD. The parameter λ of the best fitting Poisson distribution is given for the 6 low diversity subjects, together with 95% confidence intervals. Days since infection are calculated from the Poisson fit based on the value of λ using Eq. (22) with $\tau_a = 1.5$ days (Keele et al., 2008 used $\tau_a = 1.6$ days, giving slightly different estimates). In the MC method we computed the HD_l distribution for each time n and then used the Kolmogorov–Smirnov statistic to pick the distribution that best fit each patient's HD_l distribution (results shown for asynchronous model only). The mean time since the MRCA and 95% confidence intervals for the MC best fit are obtained from 10^3 simulations. The first six subjects, all of whom were characterized as having had a homogenous infection were in Fiebig stage II, implying that their estimated time from infection had a mean of 22 days and ranged between 18 and 34 days (Table 1).

for each of the homogeneous samples, and then assumed that the consensus sequence was the founder sequence s_0. We then proceeded to analyze the data in two steps: first, we used our models to infer what the expected maximum HD would be under a homogeneous infection, and used this to distinguish homogeneous infections from heterogeneous ones. Then, we applied our models to the homogeneous samples in order to infer a plausible time since the infection.

The eight early infection samples presented two distinct behaviors: either overall low diversity (max HD below 0.5%) and unimodal intersequence HD distribution, or overall high diversity (max HD above 2.5%; patients 1051 and BORI) and multimodal intersequence HD distribution (see Table 4 and Fig. 8). We hypothesized that the underlying difference between the two patterns was whether a single HIV-1 strain entered the host or if multiple distinct strains entered. In order to exclude single strain infection in the high diversity samples, we computed the maximum HD one could reasonably expect to achieve at a given point in the infection and compared this to predictions based on the Fiebig stage. To do this, we used the maximum HD probability distribution computed in the asynchronous model, Eqs. (22) and (23), with P taken from Eq. (13), using the patient specific base length N_B and sample size N_S given in Table 3. To be conservative, we then computed the combined weight in the upper 2.5% tail that would be achieved at each generation step a. The maximum HD at generation step a is less than or equal to this value 97.5% of the time. For patient 1051 (Fiebig stage III), with probability 0.975 a minimum of 241 generation steps would be needed to achieve the observed maximum HD of 71 if the infection were homogeneous. This is not compatible with Fiebig stage III, which corresponds to a maximum of 37 days since infection (Table 1). Similarly, patient BORI (Fiebig stage II) yielded 261 generation steps for a maximum HD of 77, again incompatible with the Fiebig stage. We thus conclude that the high diversity samples were from multiple strain infections, whereas the low diversity samples were compatible with a single strain infection.

An alternative way for distinguishing between homogeneous and heterogeneous infections can be obtained by looking at both the mean diversity and the variance of each sample. According to our analytical results, when the infection is truly homogeneous, the measured average diversity and its variance should be approximately equal, which is one of the basic properties of a Poisson distribution. However, since measured mean HD and variance are affected by sample size and stochastic effects in early evolution, they may differ. Thus, for a homogeneous infection we can require that the mean (diversity) and variance be located between the upper and lower 95% confidence limits computed from MC simulations based on the number of sequences sampled per patient (Fig. 9). All of the six low diversity subjects (WEAU, WITO, TRJO, SUMA, 1054, 1056) satisfy this condition, while subjects 1051 and BORI violate it. Thus we classify subjects 1051 and BORI as "heterogeneous" infections (i.e., infections initiated by two or more founder strains) and the other six subjects as "homogeneous" infections. The limitation of this classification is the assumption that the sequence population diversifies without any selection, e.g., under neutral evolution. This assumption does not hold for Fiebig stage III or higher, and this classification becomes problematic. The area outside the cone in Fig. 9 indicates

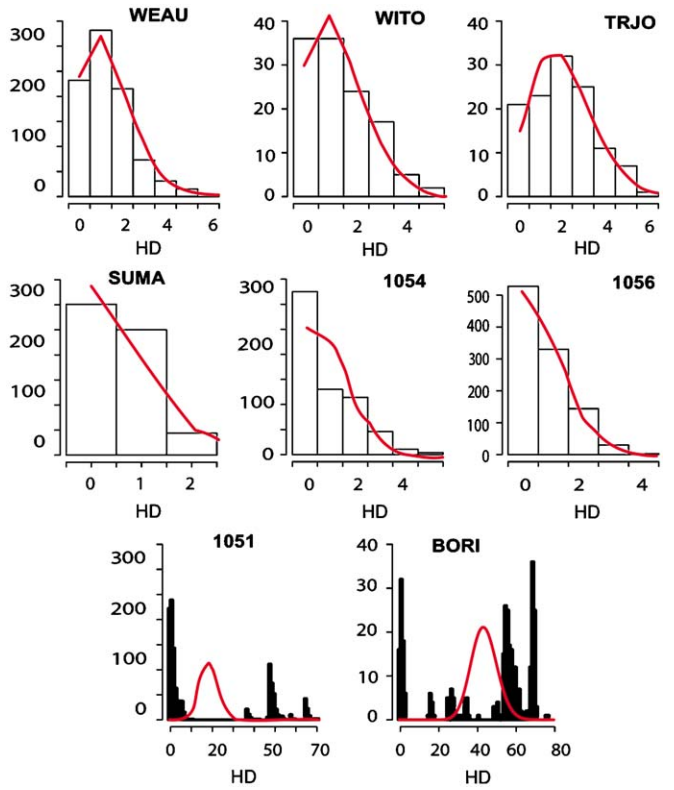

Fig. 8. Poisson fit of intersequence Hamming distance distribution. The intersequence HD distributions of eight acute HIV-1 subjects (black boxes) with the best fitting Poisson distribution (red lines) with parameter λ given by Eq. (21). Heterogeneous subjects 1051 and BORI clearly do not present a Poisson behavior. The vertical axis scales differ among patients due to difference in the number of sequences per individual. The three patients in the top row probably represent older infections than those in the middle row (Table 4) reflected in their having a greater mean HD. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

deviations from a Poisson distribution. In early Fiebig stages this suggests a heterogeneous infection, whereas in later stages it indicates either immune selection or a heterogeneous infection or both. Also due to purifying selection, samples from a heterogeneous infection could be located inside the cone in later Fiebig stages. However, to be classified as homogenous we also require that the average % diversity in these samples be less than that expected based on the upper limit of the cumulative duration for the Fiebig stage, i.e., day 34 for Fiebig stage II (Fig. 9, vertical dashed line).

The above classification was also confirmed by visual inspection through the LANL tool *Highlighter*. The visualization tool represents each sequence in the sample with a horizontal line, and places a mark whenever the sequence presents a mutation from the consensus. Fig. 9 shows two exemplary behaviors found in acute infection samples: the homogeneous sample SUMA has few randomly distributed mutations, whereas the heterogeneous sample 1051 presents a majority of sequences that are similar to the consensus and a second group of sequences that not only have many more mutations relative to the consensus, but these mutations are shared, indicative of a second lineage.

2.8. Examining neutral evolution: star-like phylogeny

During rapid exponential growth in the absence of selection, small samples of sequences are likely to have evolved from the founder sequence following a star-like phylogeny, i.e., they all coalesce at the founder (Wakeley, 2008). When this is the case, the intersequence *HD* frequency distribution, HD_l, coincides with the self-convolution of the HD_0 frequency distribution. To test whether this is the case and hence if the observed samples evolved following a star-like phylogeny, we compared the observed HD_l frequencies from those computed assuming a star-like phylogeny. Given HD_0 frequencies $X_0, X_1, \ldots, X_m$, where X_i is the number of sequences with $HD_0 = i$, we constructed the theoretical HD_l frequencies by computing the *HD* self-convolution frequencies $\overline{Y}_0, \overline{Y}_1, \ldots, \overline{Y}_n$, where $n = 2\,\mathrm{m}$, as follows:

$$\overline{Y}_k = \frac{1}{2}\sum_{i=0}^{k} X_{k-i}X_i + \frac{1}{2}\delta_{k,k-i}X_k \text{for } k = 1, \ldots, n. \tag{17}$$

The above formula is derived from the arithmetic convolution of the frequencies X_k with themselves. For example, $\overline{Y}_0 = X_0^2$, $\overline{Y}_1 = X_0X_1$, $\overline{Y}_2 = X_0X_2 + X_1^2$, and so on.

Fig. 10 compares the theoretical frequencies $\overline{Y}_0, \overline{Y}_1, \ldots, \overline{Y}_n$ (red lines), with the observed ones (the histograms). For 4 out of the 6 low diversity samples there was a perfect agreement between the calculated frequencies and the actual intersequence HD frequency. For the other two (WEAU and TRJO) a 5% and 15% difference was found, respectively. These very low differences, if any, in observed and computed frequencies suggest that all 6 samples evolved following an approximate star-like phylogeny.

2.9. Estimating time since the MRCA

We applied the synchronous and asynchronous models to characterize early homogeneous infections and estimate the time since the MRCA. We did this using both the theoretical Poisson model and the output from the MC simulations.

In a homogeneous infection we expect the intersequence HD to follow a Poisson distribution, as in Eq. (13). We used a maximum likelihood method to fit a Poisson distribution to the observed HD_l frequencies and estimate the Poisson parameter λ, which is proportional to the time or the number of generations since the most recent common ancestor. Given the vector of observed frequencies $Y = (Y_0, \ldots, Y_n)$, where Y_i is the number of sequence pairs with $HD = i$, the log likelihood function is defined as

$$LL(\lambda) = \log\left(\frac{(Y_0 + \cdots + Y_n)!}{Y_0! \ldots Y_n!}\prod_{i=0}^{n}\left[\frac{e^{-\lambda}\lambda^i}{i!}\right]^{Y_i}\right). \tag{18}$$

By minimizing the log likelihood function, we obtain

$$\lambda = \frac{\sum_{i=0}^{n} iY_i}{\sum_{i=0}^{n} Y_i}, \tag{19}$$

the mean of the intersequence Hamming distance distribution.

We then used λ to estimate the number of generation steps, n, since the most recent common ancestor (MRCA), using the following:

$$n = \begin{cases} \dfrac{\lambda}{2\varepsilon N_B} & \text{synch. model,} \\ \dfrac{\varphi}{1+\varphi}\left(\dfrac{\lambda}{\varepsilon N_B} - \dfrac{1-\varphi}{\varphi^2}\right) & \text{asynch. model,} \end{cases} \tag{20}$$

where $n = a$ in the synchronous model, and $n = \lfloor t/\tau_a \rfloor$ in the asynchronous model. Then t, the time since the MRCA, is given by $t = n\tau_s$ for the synchronous model, and $t = n\tau_a$ for the asynchronous model. As illustrated in Materials and methods, we chose values of 2 and 1.5 days for τ_s and τ_a, respectively.

We assessed the goodness of the fit using a χ^2 goodness-of-fit test statistic calculated from a singular value decomposition of the covariance matrix (see Materials and methods). In a goodness of fit test the null hypothesis is that the two distributions being

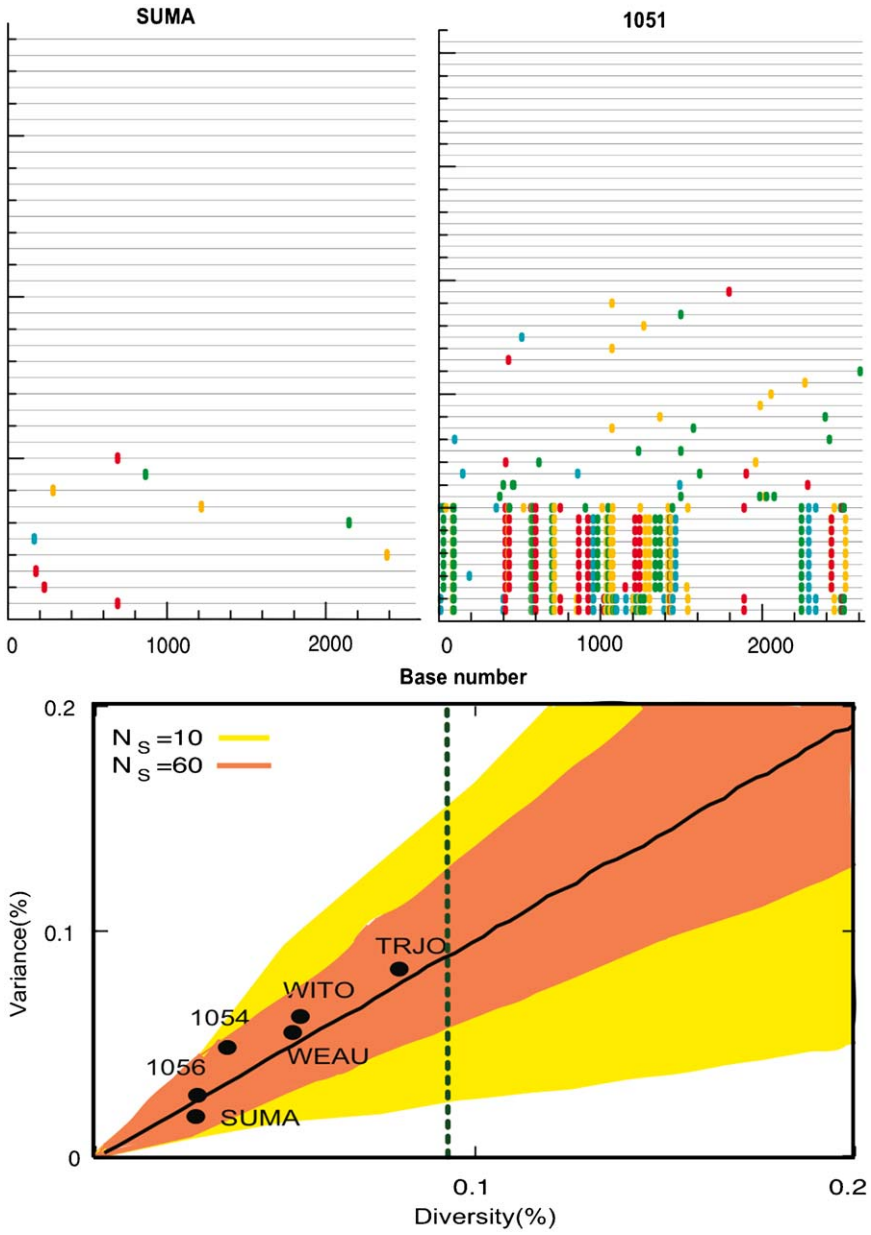

Fig. 9. Highlighter plots and formal classification diagram. The *Highlighter* plot (LANL) for subject SUMA (left) shows few and scattered mutations, whereas the analogous for subject 1051 shows frequent and aligned base substitutions (each line in the plots represents a sequence in the sample and the colored ticks represent mutations from the consensus; lines with no ticks are identical to the consensus). In the bottom panel, the diversity and the variance of the sample sequences from subjects with "homogenous" infections (i.e., infections with a single founder strain) are expected to be located within a conical area that depends on the sample size. Here we have used sample sizes of 10 and 60 to draw the yellow and orange areas, respectively, which together correspond to the 95% CI from 10^3 MC runs. The black diagonal line denotes the average relationship between diversity and variance, and the dashed vertical line denotes the average % diversity expected at day 34, which is the upper limit of the cumulative duration for Fiebig stage II (see Table 1). Samples 1051 and BORI are classified as "heterogeneous" infections since there diversities are 0.73% and 1.7% (Table 4), respectively, which places them outside this window.

tested are statistically the same, hence a low p-value rejects the null hypothesis. All 6 low-diversity samples yielded a goodness-of-fit p-value of 0.1 or higher, suggesting that the observed distribution is consistent with a Poisson.

We fixed the values of base length N_B and sample size N_S appropriate for each patient and then, under either the synchronous or the asynchronous model, we generated the HD intersequence distribution for generation steps n_1 through n_2. For example, a typical output could be generated by choosing $N_B = 2{,}600$, $N_S = 30$ and generation steps 1 through 100. At each generation step, N_S sequences were randomly sampled from it and the intersequence HD distribution of the sample was calculated. From the set of n_2-n_1+1 intersequence HD distributions generated, we used a Kolmogorov-Smirnov statistic (Casella and Berger, 1990) to pick the HD distribution that best fit the HD distribution computed from the observed sequence data. We repeated this 10^3 times. From n^*, the average of generation step of the 10^3 best-fitting distributions, we estimate the time since the MRCA, as done above for the Poisson model.

For all six low-diversity samples the time of infection could be estimated based on patients' Fiebig stage. Table 4 summarizes the estimates both from fitting the Poisson distribution and the MC output using the asynchronous model. The computed estimated time since infection estimated by each patient's Fiebig stage were within the 95% CIs predicted by the MC simulations. In contrast, the Poisson fit yielded slightly lower estimates (with respect to the given

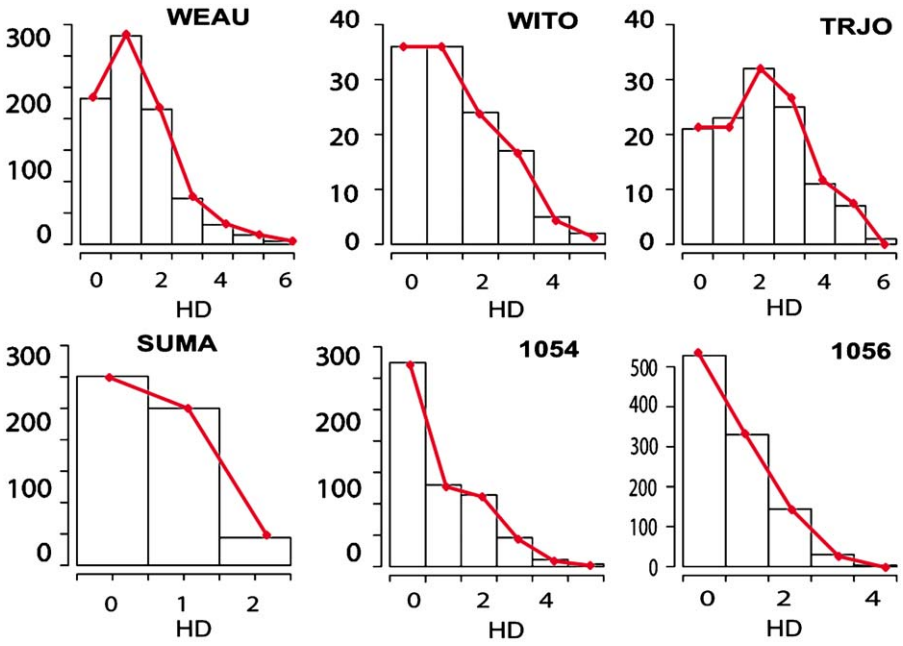

Fig. 10. Comparison between observed intersequence HD frequencies and the theoretical frequencies calculated from Eq. (19). The histograms are of the observed intersequence HD frequencies, whereas in red are the frequencies computed according to Eq. (19). The two match perfectly except for WEAU, for which an overall 5% difference was found, and TRJO, for which the overall difference was 15%. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fiebig stage) for subjects SUMA and 1056. All Poisson estimates were systematically lower than the MC ones. This is due to the fact that run-to-run variation, while accounted for in the MC fitting, is neglected when fitting the Poisson distribution. To understand this, consider for example a base length of $N_B = 2600$. From Eq. (12), under the asynchronous model, the average Hamming distance from the founder strain, after one generation step will be 0.05. In this time, only 3 infected cells are produced, and, on average there will correspondingly be 0.15 mutations. However, in any given run, we can only have a discrete number of mutations. In particular, the first three cells will be identical 85% of the time (hence a mean HD of 0), 14% of the time there will be exactly one mutation across all three sequences (hence a mean HD of 0.333), and so on. Thus, though averaging over enough runs will lead to the correct mean value, most of the runs (85%) will have less than the expected number of mutations (after the first reverse transcription), and the rest will have more. This initial difference will persist through the generations and appear as an offset in a plot of mean HD versus generations (not shown). Averaging over 1000 MC runs one finds that a sample of sequences with mean HD from the founder of 0.8 coalesce at a MRCA that is always 17 generation steps back (or 25.5 days) when fitting the Poisson distribution, but in the MC method 85% of the time the inferred number of generation steps will be higher (19 or 20 instead of 17) leading to a mean of 18.2 generation steps (or 27.3 days) Thus explaining the roughly 2 day discrepancy in the time to MRCA between these methods as shown in Table 4. Because the variation in the MC is equivalent to host-to-host variation, the MC 95% CIs are more likely to capture the true time estimates than the Poisson fits.

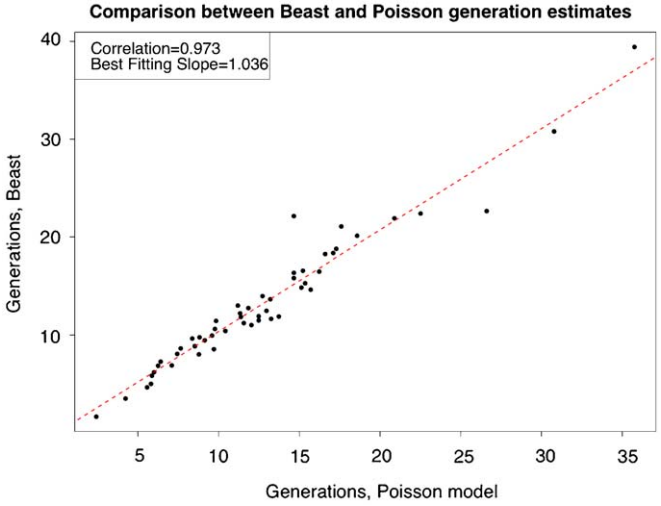

Fig. 11. Comparison between BEAST results and Poisson model results. The number of generations per sample obtained fitting the Poisson model (x-axis) and the ones obtained running the BEAST analysis (y-axis) over a set of 53 patients (Keele et al., 2008).

3. Discussion

Early HIV-1 infection tends to be characterized by a viral population with limited sequence diversity in many but not all individuals. A recent study examined sequence data obtained from 102 individuals (Keele et al., 2008), and using the methods developed in this paper characterized the infections as being consistent with a single transmitted strain in 78 cases and multiple strains in 21 cases with 3 being borderline. In order to do this classification the authors used the model of early HIV-1 infection developed in this paper.

Even though the biological conclusion that some people are infected by multiple strains and others by one strain only has already been published (Delwart et al., 2001; Derdeyn et al., 2004; Long et al., 2002; Sagar et al., 2004; Zhang et al., 1993), our goal here was to provide, in a self-contained manner, the mathematical underpinnings for the analysis of sequences obtained by single genome amplification methods.

Our goals were not novel in the sense that coalescent and Bayesian inference methods (Drummond et al., 2005, 2006; Kuhner and Smith, 2007; Kuhner et al., 1998; Rannala, 1997), when restricted to homogeneous infections, can provide the same results. In fact, comparing the estimated time to the MRCA using BEAST (Drummond and Rambaut, 2007) and using our method on the 53 patients that had a homogeneous infection and no overt enrichment for APOBEC3 mutations from Keele et al. (2008), we found the results to be similar (correlation coefficient = 0.973, best fitting slope = 1.036, see Fig. 11). However, the main difference between these methods and ours is that whereas the former simulate genealogies, under similar assumptions, our simulation models follow the entire population. While for most of the problems studied in the literature simulating genealogies is the correct approach, what one gives up in that approach is simplicity when other processes like recombination and selection need to be modeled. A forward simulation like ours, on the other hand, is usually not feasible because of the problem size, for example for studying HIV-1 at the population level, but is readily applied to the important problem of modeling the evolutionary events in early infection and characterization of viral transmission. Because of the small number of generations since the extreme bottleneck, we can follow the entire population in a simulation as well as provide analytical calculations for various quantities summarizing viral evolution. In this paper, we set up a basic framework in which we neglect recombination and selection, but it is easy to see that with our methods these pose no problems of principle, whereas the analysis involving ancestral recombination graphs (Minichiello and Durbin, 2006), or ancestral selection graphs (Sharma, 1977) or both combined would be prohibitive for this data.

For cases of a single strain infection, our model suggests that the consensus sequence, which corresponds to our best estimate of the MRCA given the Poisson model described here, is either the transmitted strain or a strain one to two mutations away from the transmitted strain (Keele et al., 2008). Identifying the transmitted/founder strain has great biological significance as one can then determine if it has particular properties that allowed it to be transmitted. Even if our identification is only accurate to within one or two base accuracy this will still provide an enormous advance as it is the virus that selectively expands, and we can use this information to search for particular attributes of the virus that allowed it to expand. For example, preliminary examination of putative transmitted viruses suggests that they may generally be more uniformly refractive to neutralization than viruses obtained at the chronic stage of infection, which show heterogeneity in susceptibility to antibodies directed at their receptor binding surfaces (Keele et al., 2008). Also there are subtype specific attributes in terms of variable loop lengths and the number of glycosylation sites (the loops appear to be shorter and less glycosylated in the transmitted viruses of A and C clade infections (Chohan et al., 2005; Derdeyn et al., 2004; Sagar et al., 2006), but not B clade transmitted viruses (Frost et al., 2005). Data is just beginning to accrue that will enable further exploration of potential viral transmission signatures and phenotypic traits, and the models described in this paper allow one to easily determine whether or not the MRCA or consensus sequence of an early sample is indeed a reasonable estimate of the transmitted virus.

Another important issue that our model addressed is the number of viruses from a single patient that one must sequence to determine if the infection is homogeneous within some confidence band and the number of sequences needed to obtain a reasonable estimate of the time to the MRCA. In order to achieve this, we computed pairwise Hamming distance frequencies of each given sample. Other types of statistical analyses that involve pairwise Hamming distances have been devised elsewhere (Gilbert et al., 2005), however, the focus there was to simply compare two different populations, not to infer time estimates. Furthermore, we showed that under the homogeneous infection hypothesis, the HD frequencies follow a star-phylogeny and Poisson distribution. This was also shown in Slatkin and Hudson (1991), where the Poisson was used to calculate a population growth rate. Here we used it to test the homogeneity of the sample, and then to get time estimates since the MRCA. The probability of misclassifying an infection as homogeneous is easily calculable from our model. For example, we showed that with a sample of 30 sequences the probability of mistakenly classifying an infection as homogeneous due to not detecting a second minor lineage ($<10\%$ of the population) is at most 5%. However, now with deep sequencing methods, such as 454 technology (Margulies et al., 2005), 100,000 or more sequences of limited length can be obtained. Such technology may indeed show examples of what we have called homogeneous infections that are not truly homogenous, and could enable us to discern if minor lineages exist at very low frequencies.

The model assumes that the viral population grows exponentially with no selection pressure and no recombination. These are reasonable assumptions to make early in the infection, before the host's immune response begins and while the infected cell population is still much smaller than the population of target cells. Further, early in infection derived from a single strain viral diversity is low and recombination may have only a small effect—as recombination between identical sequences leaves the sequences unchanged. Later in the infection, the exponential growth phase stops and selection pressure from the host environment gets established, both of which break the star-phylogeny evolution and hence our model is no longer valid.

In our mathematical models we neglect the occurrence of a mutation occurring more than once at the same site since it occurs with probability $O(a^2\varepsilon^2 N_B)$. For our values of $\varepsilon = 2.16 \times 10^{-5}$ and$N_B \sim 2600$, this probability remains below 0.006 throughout the first 100 replication cycles. After that, the model assumptions may fail because of immune pressure.

We assumed that the viral diversity evolves under a star-like phylogeny, and that all genomes coalesce at the founder strain. We used this assumption to construct the intersequence HD distribution and compute sample diversity. The probability of not coalescing at the founder strain is $O(R_0^{-m})$, where R_0 is the basic reproductive ratio and m is the number of generations away from the founder strain. So, for example, if we sample after 10 generation steps, the probability of any two randomly chosen sequences to coalesce one generation step back is $\sim R_0^{-9}$. However, the probability of coalescing 1 or 2 generations earlier than the actual founder strain is not negligible. Even though we assumed that any two sequences coalesce at the founder strain, there is in fact $<0.5\%$ chance of being three generations away from the actual founder, a 3% chance of being 2 generations away, and a 17% chance of being 1 generation away. For a set of 20–50 sequences, the probability of at least one pair coalescing after the actual founder is virtually certain, but the exponential growth of the infected cell population will severely limit the violation of star phylogeny.

It has been estimated that a free virus has a half-life of less than an hour (Ramratnam et al., 1999), whereas an infected cell has an average lifespan of $\sim$2 days (Markowitz et al., 2003; Perelson et al., 1996). Therefore, the virus present in plasma in a given sample is a reflection of the viral genomes contained in HIV-1 producing cells. Thus, as a simplification, we chose in our models to follow the infected cell population and not the virus. At each generation step, we assumed that an infected cell successfully infects R_0 other cells. In effect, we are following the integrated HIV-1 provirus rather than viral particles. The

experimental data we analyzed was obtained by sequencing virus and not provirus. On average, each cell produces as many as 5×10^4 virions (Chen et al., 2007), which makes the viral population at any given time point much larger than the infected cell population. Sampling with replacement of the provirus pool is therefore a good approximation, introducing a fractional error on the estimated probabilities of $O(10^{-4}\,N_S)$, where N_S is the number of sequences sampled. Because we never sample over 60 sequences, this is a negligible correction.

Finally, we assumed the mutation rate is constant across different lineages and along the entire genome. This assumption is violated if APOBEC3G/F causes G-to-A mutation at an enhanced rate even in sequences that are not overtly hypermutated. Keele et al. (2008) showed that overall APOBEC enriched samples follow our evolutionary model upon removing the G-to-A mutations with APOBEC3G/F signature. For the point mutation rate, we used the value 2.16×10^{-5} per base per replication cycle. Mansky and Temin (1995) estimated a mutation rate of 3.14×10^{-5} per base per replication cycle, but their estimate counted all types of mutations, including gaps. Because we exclude gaps throughout our analyses, we removed their contribution from the Mansky calculation and obtained the value 2.16×10^{-5}. The value from Mansky (1996) and Mansky and Temin (1995) was obtained *in vitro* from HeLa cells and reproduced in two different cell lines. More recent *in vitro* studies found a slightly smaller point mutation rate of 2.2×10^{-5} (including deletions and insertions) in one case (Huang and Wooley, 2005), and a larger rate of 5.4×10^{-5} in another (Gao et al., 2004). The use of a base substitution rate different from 2.16×10^{-5} would result in larger estimates for the number of days to the MRCA than obtained in Table 4, if the rate were smaller, and shorter if it were larger.

Within both the analytical and the MC model descriptions, we have envisioned two scenarios, one in which all infection events are synchronous, and one in which they are asynchronous and occur at two distinct times. Again, the advantage of the former is its simplicity. It yields a straightforward mathematical description, a population that grows exponentially in time with base R_0. However, it is biologically unrealistic that each cell would infect all other cells in at a single time. To explore how the synchronicity affects the model, we broke this assumption in the simplest possible way, e.g., by allowing infections to happen at two times. Even though this second scenario is still biologically unrealistic, nonetheless, it shows that the intersequence HD distributions do not change nature: they are still essentially Poisson distributions, and the increase in diversity is again driven by the number of RT cycles that have occurred. However, whereas in the synchronous model the increase in diversity depends on the base substitution rate ε and the generation time expressed in days, in the asynchronous model it also depends on R_0. This poses an additional difficulty in estimating time since MRCA, as studies have shown the wide variability in estimates of R_0 across patients (Little et al., 1999; Stafford et al., 2000).

For the six patients that we classified as having been infected by a single viral strain, we estimated time since the MRCA by computing the divergence (i.e., the mean Hamming distance per base pair) from the founder strain. Because we do not know the actual founder strain, we constructed the sample consensus sequence and assumed it to be identical to the MRCA. In doing this, we are implicitly assuming no bottleneck after the virus has entered the host. Should a sublineage prevail over the rest of the population after entering the host, this would cause the consensus to represent the dominant sublineage rather than the actual infecting strain; an estimate to that MRCA that is statistically less than the minimal time from infection based on Fiebig stage would point to such a scenario, and suggest selection. All these caveats factor into our time since MRCA estimates, and would similarly impact the use coalescent methods in such estimates.

We have presented a simple description of an early homogeneous infection. This restricts the validity of our model to early infections, before selection pressure can be detected and while the infected cell population is still small with respect to the target cell population. Furthermore, in some cases there may be evidence of purifying selection even early in the infection, which could explain why for some of the samples we get too early time estimates, suggesting a bottleneck in viral evolution posterior to infection. Using the *SNAP* program (LANL), we looked at the d*S*/d*N* ratios of all six low diversity patients. Four out six (1054, SUMA, TRJO, WEAU) had ratios >1, suggesting purifying selection (to be published elsewhere). Future developments of our model could allow correction for such a scenario, for example by using a modified base substitution rate.

Despite its limitations, our model successfully distinguishes homogenous from heterogeneous infections, and predicts the time evolution of divergence, diversity, maximum diversity and % sequence identity in single-strain HIV-1-infections.

4. Materials and methods

4.1. Sequence data analysis

Plasma samples were obtained from 8 subjects with acute or very recent HIV-1 subtype B infection. All subjects gave informed consent, and plasma collections were performed with institutional review board and other regulatory approvals. Blood specimens were generally collected in acid citrate dextrose and plasma separated and stored at -20 to $-70\,^{\circ}$C. To determine how the far into the acute phase the of infection the samples were taken, they were tested for HIV-1 RNA, p24 antigen, quantitative Chiron bDNA 3.0 or Roche Amplicor viral RNA assays; Coulter or Roche p24 Ag assays; Genetic Systems Anti-HIV-1/2 Plus O and Abbott AntiHIV-1/2 3rd Generation EIAs; and Genetic Systems HIV-1 Western Blot Kit. Based on these test results, subjects were staged according to the Fiebig laboratory classification system for acute and early HIV-1 infection (see Table 1).

We employed single genome amplification (SGA) of plasma HIV-1 RNA followed by direct sequence analysis of uncloned *env* amplicons (Keele et al., 2008) as described in Salazar-Gonzalez et al. (2008). The number of *env* sequences, ranging from 16 to 50, analyzed per subject, their GenBank accession numbers are given in Keele et al. (2008), and the subjects' Fiebig stages are shown in Table 1. Sequences were aligned with *GeneCutter* (LANL) and the alignment was hand-checked. We removed insertions and deletions using the *Gap Strip* tool (LANL), which removes all columns that contain one or more gaps from a nucleotide alignment.

5. Mathematical derivations

5.1. Number of mutations

We show that

$$P(\text{mutations} > N_B|a) \le O\left(\frac{a^2\varepsilon^2}{N_B}\right) \ll 1.$$

Chebyshev's inequality (Feller, 1957) states that, given a random variable X with mean μ and variance σ^2, for any positive integer k one has

$$Prob(X-\mu| \ge k\sigma) \le \frac{1}{k^2}.$$

Let X be the number of mutations. Then $\mu = a\varepsilon N_B$ and $\sigma^2 = a\varepsilon N_B(1-\varepsilon)$. We pick k such that $k\sigma + \mu = N_B + 1$, from which it follows that:

$$Prob(HD_0 \geq N_B + 1) \leq \frac{a\varepsilon(1-\varepsilon)}{N_B\left(1 + \frac{1}{N_B} - a\varepsilon\right)^2} \approx \frac{a\varepsilon(1-\varepsilon)}{N_B} + O\left(\frac{a^2\varepsilon^2}{N_B}\right) \ll 1.$$

5.2. Coalescence

To see that the probability of coalescence before generation 0 approaches zero exponentially as a and m get large, i.e., the expression in Eq. (6) is $O(R_0^{-m})$, recall that

$$x^n - 1 = (x-1)(x^{n-1} + \cdots + x + 1).$$

Therefore,

$$\frac{R_0^{a-m} - 1}{R_0^a - 1} = \frac{R_0^{a-m-1} + \cdots + R_0 + 1}{R_0^{a-1} + \cdots + R_0 + 1} = O(R_0^{a-m-1-a+1}) = O(R_0^{-m}).$$

5.3. The expected maximum HD_I

Given a sample of N_S sequences drawn from the viral population at generation a derived from a single strain infection, and a positive integer M, we examine how the N_S HD_0 values should be distributed so that the maximum intersequence Hamming distance is exactly M. Notice that $HD_I = M$ if, for any given $k \geq 0$, there is at least one sequence such that $HD_0 = k$ and another one such that $HD_0 = M - k$. Now, if $k > \lfloor M/2 \rfloor$, where $\lfloor x \rfloor$ is the floor functions defined as the maximum integer less than or equal to x, and there is more than one sequence with $HD_0 = k$, then for two such sequences their relative Hamming distance will be $HD_I = 2k > M$, assuming that for a single strain infection the MRCA is s_0. Therefore, for the maximum to be exactly M, there can be only one sequence with $HD_0 = k$, $k > \lfloor M/2 \rfloor$, and at least one with $HD_0 = M - k$, and all the others with $HD_0 \leq M - k$. Using the abbreviated notation $P_k^{(a)} = P(HD_0 = k|a) = Binom(k; aN_B, \varepsilon)$, the probability that the maximum $HD_I = M$ is thus $P_k^{(a)}$ multiplied by the probability of having all the other $N_S - 1$ sequences with $HD_0 \leq M - k$ and at least one with $HD_0 = M - k$, yielding

$$P(HD_{\max} = M|a) = P_k^{(a)}\left[\left(\sum_{i=0}^{M-k} P_i^{(a)}\right)^{N_S-1} - \left(\sum_{i=0}^{M-k-1} P_i^{(a)}\right)^{N_S-1}\right].$$

In the special case when M is even and $k = M/2$, we need at least two sequences such that $HD_0 = M/2$, which is expressed in the following:

$$\left(\sum_{i=0}^{M/2} P_i^{(a)}\right)^{N_S} - \left(\sum_{i=0}^{M/2-1} P_i^{(a)}\right)^{N_S} - N_S P_{M/2}^{(a)}\left(\sum_{i=0}^{M/2-1} P_i^{(a)}\right)^{N_S-1}. \quad (21)$$

Putting everything together, we obtain

$$P(HD_{\max} = M|a) = \sum_{k=M+1/2}^{M} N_S P_k^{(a)}\left[\left(\sum_{i=0}^{M-k} P_i^{(a)}\right)^{N_S-1} - \left(\sum_{i=0}^{M-k-1} P_i^{(a)}\right)^{N_S-1}\right], \quad (22)$$

when M is odd, and

$$P(HD_{\max} = M|a) = \left(\sum_{i=0}^{M/2} P_i^{(a)}\right)^{N_S} - \left(\sum_{i=0}^{M/2-1} P_i^{(a)}\right)^{N_S} - N_S P_{M/2}^{(a)}\left(\sum_{i=0}^{M/2-1} P_i^{(a)}\right)^{N_S-1} + \sum_{k=M/2+1}^{M} N_S P_k^{(a)}\left[\left(\sum_{i=0}^{M-k} P_i^{(a)}\right)^{N_S-1} - \left(\sum_{i=0}^{M-k-1} P_i^{(a)}\right)^{N_S-1}\right], \quad (23)$$

when M is even. Here we are assuming that $a > 0$ and that any two sequences coalesce at generation 0. Therefore, the expected maximum HD at generation a is given by

$$\overline{HD}_{\max} = \sum_{M=0}^{N_B} M \times P(HD_{\max} = M|a).$$

5.4. Asynchronous model, Eq. (8)

Of all infected cells $I(a+1,n+1)$ at time $n+1$, α $I(a,n)$ were infected by cells of age a that were "born" at time n, whereas $\gamma I(a,n-1)$ were infected by cells of age a that were born at time $n-1$. In other words, we have the following relationship:

$$I(a+1, n+1) = \alpha I(a,n) + \gamma I(a, n-1). \quad (24)$$

Using the binomial expansion

$$\binom{x+1}{y-x} = \binom{x}{y-x} + \binom{x}{y-x-1},$$

one can verify that for $\lceil n/2 \rceil \leq a \leq n$ the expression

$$I(a,n) = I_0 \alpha^a \left(\frac{\gamma}{\alpha}\right)^{n-a} \binom{a}{n-a}$$

satisfies Eq. (24) as shown below:

$$\begin{aligned}\alpha I(a,n) + \gamma I(a,n-1) &= \alpha^{a+1} I_0 \binom{a}{n-a}\left(\frac{\gamma}{\alpha}\right)^{n-a} + \gamma\alpha^a I_0 \binom{a}{n-a-1}\left(\frac{\gamma}{\alpha}\right)^{n-a-1} \\ &= \alpha^a I_0 \left(\frac{\gamma}{\alpha}\right)^{n-a}\left[\binom{a}{n-a} + \binom{a}{n-a-1}\right] \\ &= \alpha^{a+1} I_0 \left(\frac{\gamma}{\alpha}\right)^{n-a} \binom{a+1}{n-a} = I(a+1, n+1).\end{aligned}$$

5.5. Asynchronous model, Eq. (10)

Let

$$N(n) = \sum_{a=\lceil n/2 \rceil}^{n} I(a,n). \quad (25)$$

By taking the summations on both side of Eq. (24), one obtains that $N(n)$ satisfies $N(n+1) = \alpha N(n) + \gamma N(n-1)$. Let

$$\lambda_1 = \frac{\alpha}{2}(1+\varphi) \quad \text{and} \quad \lambda_2 = \frac{\alpha}{2}(1-\varphi),$$

where

$$\varphi = \sqrt{1 + 4\frac{\gamma}{\alpha^2}}.$$

Then, for any integers a and b, the function so defined: $N_{a,b}(n) = a\lambda_1^n + b\lambda_2^n$ is such that for every n it satisfies the identity $N_{a,b}(n+1) = \alpha N_{a,b}(n) + \gamma N_{a,b}(n-1)$. In fact, because λ_1 and λ_2 are the roots of the second degree equation $x^2 + ax + \gamma$, it follows that for $i = 1,2$ we have $\lambda_i^2 = \alpha\lambda_i + \gamma$. In particular, for every $n > 1$, $\lambda_i^{n+1} = \alpha\lambda_i^n + \gamma\lambda_i^{n-1}$, and as a consequence

$$\begin{aligned}\alpha N_{a,b}(n) + \gamma N_{a,b}(n-1) &= \alpha(a\lambda_1^n + b\lambda_2^n) + \gamma(a\lambda_1^{n-1} + b\lambda_2^{n-1}) \\ &= a(\alpha\lambda_1^n + \gamma\lambda_1^{n-1}) + b(\alpha\lambda_2^n + \gamma\lambda_2^{n-1}) \\ &= a\lambda_1^{n+1} + b\lambda_2^{n+1} = N_{a,b}(n+1).\end{aligned}$$

Therefore, we need to find a and b such that $N(0) = I_0$ and $N(1) = \alpha I_0$, which are the boundary conditions. This yields a linear system with solutions:

$$a = I_0 \frac{1+\varphi}{2\varphi} \quad \text{and} \quad b = -I_0 \frac{1-\varphi}{2\varphi}.$$

Hence, substituting these values into the expression of $N_{a,b}(n)$, we get

$$N(n) = I_0\left(\frac{\alpha}{2}\right)^n \frac{(1+\varphi)^{n+1} - (1-\varphi)^{n+1}}{2\varphi}.$$

As a side result, notice that if we let

$$S_n = \sum_{a=\lceil n/2\rceil}^{n} \binom{a}{n-a} (\gamma)^{n-a},$$

then it follows that:

$$S_n = \frac{(1+\varphi)^{n+1} - (1-\varphi)^{n+1}}{2^{n+1}\varphi}.$$

Finally, notice that for large values of

$$n\frac{(1+\varphi)^{n+1} - (1-\varphi)^{n+1}}{2\varphi} \approx (1+\varphi)^n,$$

and therefore

$$N(n) \approx I_0\left(\frac{\alpha}{2}(1+\varphi)\right)^n.$$

5.6. Asynchronous model Eq. (13)

In order to show that the HD_0 probability distribution is such that $P(d,n) \approx Pois(d,\lambda) + O(\varepsilon^2)$ for some suitable parameter λ, we show that up to terms in ε^2, all cumulants of this distribution are equal, which is a property that uniquely characterizes a Poisson distribution. The cumulant generating function is given by

$$\begin{aligned} K_{HD}^{(n)}(\xi) &= \operatorname{Log} M_{HD}^{(n)}(\xi) \\ &= \frac{\varepsilon N_B}{F_n} \sum_{j\geq 1} \frac{\xi^j}{j!} p_j(\varepsilon) \sum_{a=\lceil n/2\rceil}^{n} a\binom{a}{n-a}\left(\frac{\gamma}{\alpha^2}\right)^{n-a}, \end{aligned} \tag{26}$$

where $M_{HD}^{(n)}(\xi)$ is the moment generating function

$$M_{HD}^{(n)}(\xi) = \frac{1}{F_n} \sum_{a=\lceil n/2\rceil}^{n} \binom{a}{n-a}\binom{aN_B}{d}\left(\frac{\gamma}{\alpha^2}\right)^{n-a} [\varepsilon e^{\xi} + (1-\varepsilon)]^{aN_B}.$$

In Eq. (26), $p_j(\varepsilon)$ is a polynomial in ε of degree $j-1$ such that $p_j(0) = 1$ for all j's. Let

$$X(n) = \sum_{a=\lceil n/2\rceil}^{n} a\binom{a}{n-a}\left(\frac{\gamma}{\alpha^2}\right)^{n-a}.$$

One can show that for sufficiently large n, the following holds:

$$\frac{X(n)}{F_n} \approx n\frac{1+\varphi}{2\varphi} + \frac{1-\varphi}{2\varphi^2}.$$

This can be shown by first noticing that

$$\frac{F_{n+1}}{F_n} \approx \frac{1+\varphi}{2} + O(e^{-n})$$

because of the identity

$$\begin{aligned} \frac{F_n}{F_{n-1}} &= \frac{1}{2}\frac{(1+\varphi)^{n+1} - (1-\varphi)^{n+1}}{(1+\varphi)^n - (1-\varphi)^n} \\ &= \frac{1}{2}\frac{1+\varphi - (1-\varphi)\left(\frac{1-\varphi}{1+\varphi}\right)^n}{1 - \left(\frac{1-\varphi}{1+\varphi}\right)^n} \end{aligned} \tag{27}$$

and since

$$\theta = \frac{1-\varphi}{1+\varphi} < 1,$$

it follows that:

$$\frac{F_n}{F_{n-1}} \approx \frac{1+\varphi}{2} + \theta^n\left(\frac{1+\varphi}{2}\right) + O(\theta^{2n}).$$

Given

$$X(n) = \sum_{a=\lceil n/2\rceil}^{n} a\binom{a}{n-a}\left(\frac{\gamma}{\alpha^2}\right)^{n-a}$$

for $n \gg 1$, we have

$$\frac{X(n)}{F_n} \approx n\frac{1+\varphi}{2\varphi} + \frac{1-\varphi}{2\varphi^2}.$$

In fact, let

$$x = \frac{\gamma}{\alpha^2} = \frac{\varphi^2 - 1}{4}.$$

Then

$$\begin{aligned} F_n + \left(\frac{\varphi^2-1}{4}\right)\left(\frac{n\varphi+1}{\varphi^2}\right)F_{n-1} &= \frac{\partial}{\partial x}[xF_n] = \frac{\partial}{\partial x}\left[\sum \binom{a}{n-a} x^{n-a+1}\right] \\ &= (n+1)F_n - \sum a\binom{a}{n-a}x^{n-a}. \end{aligned}$$

As a consequence,

$$X(n) = nF_n - n\frac{\varphi^2-1}{4\varphi}F_{n-1} - \frac{\varphi^2-1}{4\varphi^2}F_{n-1}.$$

Dividing both sides by F_n completes the proof.

Given this notation, we can write

$$K_{HD}^{(n)}(\xi) = \varepsilon N_B\left(\sum_{j\geq 1}\frac{\xi^j}{j!}p_j(\varepsilon)\right)\frac{X(n)}{F_n}.$$

As a consequence, for large n, we can approximate the m-th cumulant as

$$\begin{aligned} \kappa_m^{(n)} &\approx \varepsilon N_B p_m(\varepsilon)\left(n\frac{1+\varphi}{2\varphi} + \frac{1-\varphi}{2\varphi^2}\right) \\ &\approx \left(n\frac{1+\varphi}{2\varphi} + \frac{1-\varphi}{2\varphi^2}\right)\varepsilon N_B + O(\varepsilon^2 N_B), \end{aligned}$$

and up to terms in $O(\varepsilon^2 N_B)$, all cumulants are equal (the m-th cumulant does not depend on m but only on n), which proves that the distribution is approximately Poisson.

5.7. Generation times τ_s and τ_a

The times at which cells are infected differ in the synchronous and asynchronous models. However, the growth rate of the infected cell population in the two models should both be chosen equal to the best estimate of the actual growth rate. Setting these growth rates equal will lead to a relationship between τ_s and τ_a. Current estimates for the generation time are roughly two days (Markowitz et al., 2003). Thus, lets assume $\tau_s = 2$ days. In order to derive the duration of τ_a, notice that $Log_{R_0}(N(n))$ is linear in n for sufficiently large n, where $N(n)$ is the quantity defined in Eq. (25). In particular, this quantity grows linearly with a rate per τ_a given by

$$\rho_a = \operatorname{Log}_{R_0}\frac{N(n+1)}{N(n)} \approx \operatorname{Log}_{R_0}\left(\alpha\frac{1+\varphi}{2}\right).$$

In the synchronous model, at any time $t = a$, the total number of newly infected cells is $N(n) = I_0 R_0^n$. Therefore, the synchronous

growth rate ρ_s in the logarithmic scale is 1, i.e.,

$$\rho_s = \mathrm{Log}_{R_0} \frac{N(n+1)}{N(n)} = \mathrm{Log}_{R_0} \frac{R_0^{n+1}}{R_0^n} = 1.$$

If the mean generation time is 2 days, it follows that the average growth rate is 0.5/day. Comparing the rate of increase per day between the two models, we get

$$\rho_a \frac{\Delta\tau_a}{\Delta days} = \rho_s \frac{\Delta\tau_s}{\Delta days} = \frac{1}{2},$$

where $\tau_s = 2$ days. Hence

$$\tau_a = 2\rho_a \approx 2\,\mathrm{Log}_{R_0}\left(\alpha\frac{1+\varphi}{2}\right).$$

In particular, when

$$\alpha = \gamma = \frac{R_0}{2},$$

$\tau_a = 1.5$ days and the total number of newly infected cells at time $t = n\tau_a$ is given by

$$N(n) = I_0\left(\frac{R_0}{4}\right)^n \frac{(1+\varphi)^{n+1} - (1-\varphi)^{n+1}}{2\varphi},$$

where $\varphi = \sqrt{1 + 8/R_0}$. We could also assume given the data that the average generation time in an asynchronous model is 2 days and then use the procedure given above to deduce a corresponding τ_s. In using a τ_a of 1.5 days half the infected cells are implicitly assumed to live for 3 days. While the average measured lifetime is 2 days the variance around the mean is not known and hence it is unclear if this value of τ_a is realistic.

5.8. Chi square test for dependent data cells

Define

$$\chi^2 = (Y - E(Y))^t \Sigma^{-1} (Y - E(Y)), \tag{28}$$

where Y is the vector of intersequence HD frequencies of length $n = \max(HD)$, E denotes the expectation operator $E(Y) = (E(Y_0), \ldots, E(Y_n))$, and Σ^{-1} is calculated by inverting the covariance matrix $\Sigma_{i,j} = \sigma_{ij} = E(Y_i Y_j) - E(Y_i)E(Y_j)$ on its non-singular space using Singular Value Decomposition. This is motivated by the non-independence of our observations, since they are drawn from a convolution of two identical Poisson distributions with mean λ. In particular

$$E(Y_k) = c^2 2^{k-1} e^{-2\lambda} \frac{\lambda^k}{k!},$$

where c is a normalization constant and $k = 0,\ldots,n$. Since the convolution operation imposes $\lfloor n/2 \rfloor + 1$ linear constraints on Σ's columns, the degrees of freedom of the χ^2 statistic are $df \leq \lfloor n/2 \rfloor - 1$.

5.9. Monte Carlo simulations

The number of bases that undergo mutation is determined by the mutation rate ε. However, once a base is chosen to be mutated a matrix of base transitions is then used to determine which base it is mutated into. This same method is used in both the synchronous and asynchronous models. The base frequencies and substitution matrix employed in the MC simulations were obtained from a general time reversible (GTR) evolutionary model (Yang, 1994) optimized in conjunction with a maximum likelihood phylogenetic tree. Relative base substitution frequencies are listed in Table 5; these were based on the GTR probability matrix (Korber et al., 2000) and a maximum likelihood estimate of rate variation per site, and indicate the probability of change of one nucleotide to another per unit time, assuming a branch length of 1 (there is extensive change with a branch length of 1). In the MC simulation the equilibrium base frequencies for this model were $A = 0.373$, $G = 0.241$, $C = 0.193$, and $T = 0.192$. The transition matrix was computed from Table 5 by setting the diagonal elements to zero and renormalizing each row.

Table 5
Transition matrix.

	A	G	C	T
A	0.5450	0.2475	0.1137	0.0937
G	0.3836	0.4357	0.0891	0.0915
C	0.2192	0.1108	0.4274	0.2425
T	0.1823	0.1149	0.2448	0.4581

Rows are "from", columns are "to".

The maximum likelihood tree used to derive this matrix in Table 5 included 445 HIV-1 clade B Envelope sequences from the Los Alamos database (LANL) combined with sequences from Keele et al. (2008) (code described in Korber et al., 2000). While this model was used for the simulations presented in this paper, the use of the HIV-1 specific GTR model turned out to have little impact on the simulation results; a neutral model with an even base distribution and equal substitution rates yielded essentially the same results for the simulations (data not shown).

Acknowledgments

Portions of this work were done under the auspices of the US Department of Energy under Contract DE-AC52-06NA25396 and supported in part by the Center for HIV/AIDS Vaccine Immunology (AI67854), the Bill & Melinda Gates Foundation Grand Challenges Program (37874), the University of Alabama at Birmingham Center for AIDS Research (AI27767), the University of Rochester Developmental Center for AIDS Research (P30-AI078498) and NIH Grants AI083115, AI028433, and RR06555. We thank Marcus Daniels for technical assistance.

References

Abrahams, M.R., Anderson, J.A., Giorgi, E.E., Seoighe, C., Mlisana, K., Ping, L.H., Athreya, G.S., Treurnicht, F.K., Keele, B.F., Wood, N., Salazar-Gonzalez, J.F., Bhattacharya, T., Chu, H., Hoffman, I., Galvin, S., Mapanje, C., Kazembe, P., Thebus, R., Fiscus, S., Hide, W., Cohen, M.S., Karim, S.A., Haynes, B.F., Shaw, G.M., Hahn, B.H., Korber, B.T., Swanstrom, R., Williamson, C., 2009. Quantitating the multiplicity of infection with human immunodeficiency virus type 1 subtype C reveals a non-Poisson distribution of transmitted variants. J. Virol. 83, 3556–3567.

Achaz, G., Palmer, S., Kearney, M., Maldarelli, F., Mellors, J.W., Coffin, J.M., Wakeley, J., 2004. A robust measure of HIV-1 population turnover within chronically infected individuals. Mol. Biol. Evol. 21, 1902–1912.

Bourara, K., Liegler, T.J., Grant, R.M., 2007. Target cell APOBEC3C can induce limited G-to-A mutation in HIV-1. PLoS Pathog. 3, 1477–1485.

Casella, G., Berger, R.L., 1990. Statistical Inference. Brooks/Cole Publ. Co., Pacific Grove, CA.

Chen, H.Y., Di Mascio, M., Perelson, A.S., Ho, D.D., Zhang, L., 2007. Determination of virus burst size in vivo using a single-cycle SIV in rhesus macaques. Proc. Natl. Acad. Sci. USA 104, 19079–19084.

Chohan, B., Lang, D., Sagar, M., Korber, B., Lavreys, L., Richardson, B., Overbaugh, J., 2005. Selection for human immunodeficiency virus type 1 envelope glycosylation variants with shorter V1–V2 loop sequences occurs during transmission of certain genetic subtypes and may impact viral RNA levels. J. Virol. 79, 6528–6531.

Clark, S.J., Saag, M.S., Decker, W.D., Campbell-Hill, S., Roberson, J.L., Veldkamp, P.J., Kappes, J.C., Hahn, B.H., Shaw, G.M., 1991. High titers of cytopathic virus in plasma of patients with symptomatic primary HIV-1 infection. N. Engl. J. Med. 324, 954–960.

Delwart, E.L., Magierowska, M., Royz, M., Foley, B., Peddada, L., Smith, R., Heldebrand, C., Conrad, A., Busch, M.P., 2001. Homogeneous quasispecies in

16 out of 17 individuals during very early HIV-1 primary infection. AIDS 15, 1–7.
Derdeyn, C.A., Decker, J.M., Bibollet-Ruche, F., Mokili, J.L., Muldoon, M., Denham, S.A., Heil, M.L., Kasolo, F., Musonda, R., Hahn, B.H., Shaw, G.M., Korber, B.T., Allen, S., Hunter, E., 2004. Envelope-constrained neutralization-sensitive HIV-1 after heterosexual transmission. Science 303, 2019–2022.
Dickover, R., Garratty, E., Yusim, K., Miller, C., Korber, B., Bryson, Y., 2006. Role of maternal autologous neutralizing antibody in selective perinatal transmission of human immunodeficiency virus type 1 escape variants. J. Virol. 80, 6525–6533.
Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214.
Drummond, A.J., Rambaut, A., Shapiro, B., Pybus, O.G., 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. 22, 1185–1192.
Drummond, A.J., Ho, S.Y., Phillips, M.J., Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4, e88.
Edwards, C.T., Holmes, E.C., Wilson, D.J., Viscidi, R.P., Abrams, E.J., Phillips, R.E., Drummond, A.J., 2006. Population genetic estimation of the loss of genetic diversity during horizontal transmission of HIV-1. BMC Evol. Biol. 6, 28.
Feller, W., 1957. An Introduction to Probability Theory and its Applications. John Wiley and Sons, Inc., New York.
Fiebig, E.W., Heldebrant, C.M., Smith, R.I., Conrad, A.J., Delwart, E.L., Busch, M.P., 2005. Intermittent low-level viremia in very early primary HIV-1 infection. J. Acquir. Immune Defic. Syndr. 39, 133–137.
Fiebig, E.W., Wright, D.J., Rawal, B.D., Garrett, P.E., Schumacher, R.T., Peddada, L., Heldebrant, C., Smith, R., Conrad, A., Kleinman, S.H., Busch, M.P., 2003. Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection. AIDS 17, 1871–1879.
Frost, S.D., Liu, Y., Pond, S.L., Chappey, C., Wrin, T., Petropoulos, C.J., Little, S.J., Richman, D.D., 2005. Characterization of human immunodeficiency virus type 1 (HIV-1) envelope variation and neutralizing antibody responses during transmission of HIV-1 subtype B. J. Virol. 79, 6523–6527.
Gaines, H., von Sydow, M., Pehrson, P.O., Lundbegh, P., 1988. Clinical picture of primary HIV infection presenting as a glandular-fever-like illness. BMJ 297, 1363–1368.
Gao, F., Chen, Y., Levy, D.N., Conway, J.A., Kepler, T.B., Hui, H., 2004. Unselected mutations in the human immunodeficiency virus type 1 genome are mostly nonsynonymous and often deleterious. J. Virol. 78, 2426–2433.
Gilbert, P.B., Rossini, A.J., Shankarappa, R., 2005. Two-sample tests for comparing intra-individual genetic sequence diversity between populations. Biometrics 61, 106–117.
Harris, R.S., Liddament, M.T., 2004. Retroviral restriction by APOBEC proteins. Nat. Rev. Immunol. 4, 868–877.
Heffernan, J.M., Wahl, L.M., 2005. Monte Carlo estimates of natural variation in HIV infection. J. Theor. Biol. 236, 137–153.
Huang, K.J., Wooley, D.P., 2005. A new cell-based assay for measuring the forward mutation rate of HIV-1. J. Virol. Methods 124, 95–104.
Kamina, A., Makuch, R.W., Zhao, H., 2001. A stochastic modeling of early HIV-1 population dynamics. Math. Biosci. 170, 187–198.
Keele, B., Li, H., Learn, G.H., Hraber, P.T., Giorgi, E.E., Grayson, T., Sun, C., Chen, Y., Yeh, W.W., Letvin, N.L., Mascola, J.R., Nabel, G.J., Hayens, B.F., Bhattacharya, T., Perelson, A.S., Korber, B.T., Hahn, B.H., Shaw, G.S., 2009. Low dose rectal inoculation of rhesus macaques by SIVsmE660 or SIVmac251 recapitulates human mucosal infection by HIV-1. J. Exp. Med. 206, 1117–1134.
Keele, B.F., Giorgi, E.E., Salazar-Gonzalez, J.F., Decker, J.M., Pham, K.T., Salazar, M.G., Sun, C., Grayson, T., Wang, S., Li, H., Wei, X., Jiang, C., Kirchherr, J.L., Gao, F., Anderson, J.A., Ping, L.H., Swanstrom, R., Tomaras, G.D., Blattner, W.A., Goepfert, P.A., Kilby, J.M., Saag, M.S., Delwart, E.L., Busch, M.P., Cohen, M.S., Montefiori, D.C., Haynes, B.F., Gaschen, B., Athreya, G.S., Lee, H.Y., Wood, N., Seoighe, C., Perelson, A.S., Bhattacharya, T., Korber, B.T., Hahn, B.H., Shaw, G.M., 2008. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proc. Natl. Acad. Sci. USA 105, 7552–7557.
Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B.H., Wolinsky, S., Bhattacharya, T., 2000. Timing the ancestor of the HIV-1 pandemic strains. Science 288, 1789–1796.
Kuhner, M.K., Smith, L.P., 2007. Comparing likelihood and Bayesian coalescent estimation of population parameters. Genetics 175, 155–165.
Kuhner, M.K., Yamato, J., Felsenstein, J., 1998. Maximum likelihood estimation of population growth rates based on the coalescent. Genetics 149, 429–434.
LANL, Hypermut. 〈http://www.hiv.lanl.gov/content/sequence/HYPERMUT/hypermut.html〉.
LANL, Consensus Maker. 〈http://www.hiv.lanl.gov/content/sequence/CONSENSUS/consensus.html〉.
LANL, Highlighter. 〈http://www.hiv.lanl.gov/content/sequence/HIGHLIGHT/highlighter.html〉.
LANL, SNAP. 〈http://www.hiv.lanl.gov/content/sequence/SNAP/SNAP.html〉.
LANL, GeneCutter. 〈http://www.hiv.lanl.gov/content/sequence/GENE_CUTTER/cutter.html〉.
LANL, Los Alamos Database. 〈http://www.hiv.lanl.gov/content/index〉.
LANL, Gap Strip. 〈http://www.hiv.lanl.gov/content/sequence/GAPSTREEZE/strip_ready.html〉.
Lee, H.Y., Perelson, A.S., Park, S.C., Leitner, T., 2008. Dynamic correlation between intrahost HIV-1 quasispecies evolution and disease progression. PLoS Comput. Biol. 4, e100240.
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.
Lindback, S., Karlsson, A.C., Mittler, J., Blaxhult, A., Carlsson, M., Briheim, G., Sonnerborg, A., Gaines, H., 2000a. Viral dynamics in primary HIV-1 infection. Karolinska Institute Primary HIV Infection Study Group. AIDS 14, 2283–2291.
Lindback, S., Thorstensson, R., Karlsson, A.C., von Sydow, M., Flamholc, L., Blaxhult, A., Sonnerborg, A., Biberfeld, G., Gaines, H., 2000b. Diagnosis of primary HIV-1 infection and duration of follow-up after HIV exposure. Karolinska Institute Primary HIV Infection Study Group. AIDS 14, 2333–2339.
Little, S.J., McLean, A.R., Spina, C.A., Richman, D.D., Havlir, D.V., 1999. Viral dynamics of acute HIV-1 infection. J. Exp. Med. 190, 841–850.
Long, E.M., Rainwater, S.M., Lavreys, L., Mandaliya, K., Overbaugh, J., 2002. HIV type 1 variants transmitted to women in Kenya require the CCR5 coreceptor for entry, regardless of the genetic complexity of the infecting virus. AIDS Res. Hum. Retroviruses 18, 567–576.
Mansky, L.M., 1996. Forward mutation rate of human immunodeficiency virus type 1 in a T lymphoid cell line. AIDS Res. Hum. Retroviruses 12, 307–314.
Mansky, L.M., Temin, H.M., 1995. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J. Virol. 69, 5087–5094.
Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., Rothberg, J.M., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380.
Markowitz, M., Louie, M., Hurley, A., Sun, E., Di Mascio, M., Perelson, A.S., Ho, D.D., 2003. 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. 77, 5037–5038.
Minichiello, M.J., Durbin, R., 2006. Mapping trait loci by use of inferred ancestral recombination graphs. Am. J. Hum. Genet. 79, 910–922.
Nowak, M.A., Lloyd, A.L., Vasquez, G.M., Wiltrout, T.A., Wahl, L.M., Bischofberger, N., Williams, J., Kinter, A., Fauci, A.S., Hirsch, V.M., Lifson, J.D., 1997. Viral dynamics of primary viremia and antiretroviral therapy in simian immunodeficiency virus infection. J. Virol. 71, 7518–7525.
Painter, S.L., Biek, R., Holley, D.C., Poss, M., 2003. Envelope variants from women recently infected with clade A human immunodeficiency virus type 1 confer distinct phenotypes that are discerned by competition and neutralization experiments. J. Virol. 77, 8448–8461.
Perelson, A.S., Neumann, A.U., Markowitz, M., Leonard, J.M., Ho, D.D., 1996. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science 271, 1582–1586.
Ramratnam, B., Bonhoeffer, S., Binley, J., Hurley, A., Zhang, L., Mittler, J.E., Markowitz, M., Moore, J.P., Perelson, A.S., Ho, D.D., 1999. Rapid production and clearance of HIV-1 and hepatitis C virus assessed by large volume plasma apheresis. Lancet 354, 1782–1785.
Rannala, B., 1997. Gene genealogy in a population of variable size. Heredity 78 (Pt 4), 417–423.
Ribeiro, R.M., Bonhoeffer, S., 1999. A stochastic model for primary HIV infection: optimal timing of therapy. AIDS 13, 351–357.
Richman, D.D., Wrin, T., Little, S.J., Petropoulos, C.J., 2003. Rapid evolution of the neutralizing antibody response to HIV type 1 infection. Proc. Natl. Acad. Sci. USA 100, 4144–4149.
Ritola, K., Pilcher, C.D., Fiscus, S.A., Hoffman, N.G., Nelson, J.A., Kitrinos, K.M., Hicks, C.B., Eron Jr., J.J., Swanstrom, R., 2004. Multiple V1/V2 env variants are frequently present during primary infection with human immunodeficiency virus type 1. J. Virol. 78, 11208–11218.
Ruskin, H.J., Pandey, R.B., Liu, Y., 2002. Viral load and stochastic mutation in a Monte Carlo simulation of HIV. Physica A 293, 315–323.
Sagar, M., Wu, X., Lee, S., Overbaugh, J., 2006. Human immunodeficiency virus type 1 V1–V2 envelope loop sequences expand and add glycosylation sites over the course of infection, and these modifications affect antibody neutralization sensitivity. J. Virol. 80, 9586–9598.
Sagar, M., Kirkegaard, E., Long, E.M., Celum, C., Buchbinder, S., Daar, E.S., Overbaugh, J., 2004. Human immunodeficiency virus type 1 (HIV-1) diversity at time of infection is not restricted to certain risk groups or specific HIV-1 subtypes. J. Virol. 78, 7279–7283.
Salazar-Gonzalez, J.F., Bailes, E., Pham, K.T., Salazar, M.G., Guffey, M.B., Keele, B.F., Derdeyn, C.A., Farmer, P., Hunter, E., Allen, S., Manigart, O., Mulenga, J., Anderson, J.A., Swanstrom, R., Haynes, B.F., Athreya, G.S., Korber, B.T., Sharp, P.M., Shaw, G.M., Hahn, B.H., 2008. Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. J. Virol. 82, 3952–3970.
Schacker, T., Collier, A.C., Hughes, J., Shea, T., Corey, L., 1996. Clinical and epidemiologic features of primary HIV infection. Ann. Intern. Med. 125, 257–264.
Shankarappa, R., Margolick, J.B., Gange, S.J., Rodrigo, A.G., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C.R., Learn, G.H., He, X., Huang, X.L., Mullins, J.I., 1999.

Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J. Virol. 73, 10489–10502.
Sharma, J.C., 1977. Inheritance of defective extremities. Indian J. Med. Res. 66, 120–126.
Simon, V., Zennou, V., Murray, D., Huang, Y., Ho, D.D., Bieniasz, P.D., 2005. Natural variation in Vif: differential impact on APOBEC3G/3F and a potential role in HIV-1 diversification. PLoS Pathog. 1, e6.
Slatkin, M., Hudson, R.R., 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129, 555–562.
Stafford, M.A., Corey, L., Cao, Y., Daar, E.S., Ho, D.D., Perelson, A.S., 2000. Modeling plasma virus concentration during primary HIV infection. J. Theor. Biol. 203, 285–301.
Tan, W.Y., Wu, H., 1998. Stochastic modeling of the dynamics of CD4+ T-cell infection by HIV and some Monte Carlo studies. Math. Biosci. 147, 173–205.
Taylor, J.R., 1982. An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements. University Science Books, Mill Valley, CA.
Tuckwell, H.C., Le Corfec, E., 1998. A stochastic model for early HIV-1 population dynamics. J. Theor. Biol. 195, 451–463.
Vernazza, P.L., Eron, J.J., Fiscus, S.A., Cohen, M.S., 1999. Sexual transmission of HIV: infectiousness and prevention. AIDS 13, 155–166.
Wakeley, J., 2008. Coalescent Theory, An Introduction. Roberts & Co.
Wei, X., Decker, J.M., Wang, S., Hui, H., Kappes, J.C., Wu, X., Salazar-Gonzalez, J.F., Salazar, M.G., Kilby, J.M., Saag, M.S., Komarova, N.L., Nowak, M.A., Hahn, B.H., Kwong, P.D., Shaw, G.M., 2003. Antibody neutralization and escape by HIV-1. Nature 422, 307–312.
Wolinsky, S.M., Wike, C.M., Korber, B.T., Hutto, C., Parks, W.P., Rosenblum, L.L., Kunstman, K.J., Furtado, M.R., Munoz, J.L., 1992. Selective transmission of human immunodeficiency virus type-1 variants from mothers to infants. Science 255, 1134–1137.
Wolinsky, S.M., Korber, B.T., Neumann, A.U., Daniels, M., Kunstman, K.J., Whetsell, A.J., Furtado, M.R., Cao, Y., Ho, D.D., Safrit, J.T., 1996. Adaptive evolution of human immunodeficiency virus-type 1 during the natural course of infection. Science 272, 537–542.
Yang, Z., 1994. Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39, 105–111.
Zhang, L.Q., MacKenzie, P., Cleland, A., Holmes, E.C., Brown, A.J., Simmonds, P., 1993. Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection. J. Virol. 67, 3345–3356.