BMC Evolutionary Biology

BioMed Central

Research article

Open Access

Selection dramatically reduces effective population size in HIV-1 infection

Yi Liu* and John E Mittler

Address: Department of Microbiology, University of Washington School of Medicine, Seattle, Washington, 98195, USA

Email: Yi Liu* - yiliu197@u.washington.edu; John E Mittler - jmittler@u.washington.edu

* Corresponding author

Published: 3 May 2008

BMC Evolutionary Biology 2008, **8**:133 doi:10.1186/1471-2148-8-133

Received: 13 December 2007

Accepted: 3 May 2008

This article is available from: http://www.biomedcentral.com/1471-2148/8/133

Abstract

Background: In HIV-1 evolution, a 100–100,000 fold discrepancy between census size and effective population size (N_e) has been noted. Although it is well known that selection can reduce N_e, high *in vivo* mutation and recombination rates complicate attempts to quantify the effects of selection on HIV-1 effective size.

Results: We use the inbreeding coefficient and the variance in allele frequency at a linked neutral locus to estimate the reduction in N_e due to selection in the presence of mutation and recombination. With biologically realistic mutation rates, the reduction in N_e due to selection is determined by the strength of selection, i.e., the stronger the selection, the greater the reduction. However, the dependence of N_e on selection can break down if recombination rates are very high (e.g., $r \geq 0.1$). With biologically likely recombination rates, our model suggests that recurrent selective sweeps similar to those observed *in vivo* can reduce within-host HIV-1 effective population sizes by a factor of 300 or more.

Conclusion: Although other factors, such as unequal viral reproduction rates and limited migration between tissue compartments contribute to reductions in N_e, our model suggests that recurrent selection plays a significant role in reducing HIV-1 effective population sizes *in vivo*.

Background

The effective population size, N_e, is defined as the size of an idealized population that has the same population genetics properties (generally those properties that measure the magnitude of random genetic drift) as the actual population. Most studies have estimated the within-host N_e for HIV-1 during chronic infection to be $\sim 10^3$ [1-5], though one study estimated N_e to be between 10^5 and 5×10^5 [6]. Even the highest of these estimates is about two orders of magnitude lower than the number of productively infected cells, estimated to be on the order of 10^7 to 10^8 cells [7]. Explanations for low N_e values include unequal viral reproduction rates [2-5,8], structured populations [8-12], and recurrent selection [2-5,8]. The possibility that recurring selection may be reducing viral diversity is unsettling because most of the computational models used to estimate N_e assume neutral evolution.

During a selective sweep of a favorable allele, any neutral alleles linked to the selected allele will rise in frequency and become overrepresented in the population. This process, called "hitchhiking", can reduce neutral diversity more than random genetic drift and therefore reduce N_e [13]. Although selection has been acknowledged as a possible explanation for the low within-host effective population size during chronic HIV-1 infection [3,12], high

mutation [14,15] and recombination rates [16-20] complicate attempts to study the effects of selection on HIV-1 *in vivo*. To address these issues, we extended a classic "inbreeding coefficient" method [21-23] to derive recurrence equations that account for the combined effects of selection, mutation, and recombination. We then used these equations to quantify the effects of selection on effective size using parameters relevant to HIV-1 evolution *in vivo*.

Results and Discussion

Overview of the genetic model

Our model follows the basic Wright-Fisher assumptions of a single haploid population of constant size with no subdivision or migration, non-overlapping generations, and random sampling of offspring each generation. We calculated N_e in terms of the inbreeding effective size, which is based on the change of the average inbreeding coefficient (F) at a neutral locus (L) that is linked to a locus (S) that is under selection. The inbreeding coefficient is defined as the probability that two individuals are identical by descent (which means they are identical and have a common ancestor). Therefore, for the neutral locus L, two individuals are identical by descent if they are derived from a common ancestor and are identical at locus L, regardless of the status of locus S. Our approach to estimating N_e was to determine changes in the inbreeding coefficient at the neutral locus in the presence and absence of selection and recombination. The effective population size was defined as the size of the neutral population that gave changes in the inbreeding coefficient that were equal to those observed in the presence of selection and recombination.

As shown in Figure 1A, in the absence of recombination, an offspring can be derived from a parent in the previous generation with either allele *a* or *A* at locus *S*. An offspring with allele *a* can be derived by two pathways: from a parent with allele *a* (without mutation) or a parent with allele *A* (with an *A* to *a* mutation). An offspring with allele *A* can be derived by two similar pathways. Therefore, F_t (the value of F at time t) will be the sum of the probability that two offspring are derived from a certain combination of parents (both with allele *A*, both with allele *a*, and one with allele *A* and the other with allele *a*) times the probability that the offspring are identical by descent at locus *L* (see Appendix).

In the presence of recombination, loci *L* and *S* can be derived from different parents (Figure 1B). An offspring with allele *a* or *A* at locus *S* can be derived from one or more parents in the previous generation by the four pathways illustrated in Figure 1B. As above, F_t will be the sum of the probability that the two offspring are derived from a certain combination of pathways (both having locus *S* from parents with allele *A*, both having locus *S* from parents with allele *a*, one having locus *S* from a parent with allele *a* and the other having locus *S* from a parent with allele *A*) times the probability that the offspring derived from these pathways are identical by descent at locus *L* (see APPENDIX).

Effect of selection on effective population size

We used the ratio N/N_e to summarize the reduction in N_e due to selection from the start of selection at $t = 0$ until $t = t_{nearlyfixed}$ [the time when the frequency of the advantageous allele reaches $(N\text{-}1)/N$]. This last approximation is helpful because fixation time is asymptotic, with the advantageous allele never reaching 100% in a deterministic model.

In the absence of mutation, the reduction in N_e due to selection was most strongly affected by the initial frequency of the advantageous allele, A_0 (Figure 2A). In the presence of mutation, the reduction in N_e due to selection was most sensitive to the selective advantage, s, of the advantageous allele (Figure 2B). Indeed, for a homogeneous population of $N = 10^7$, the N/N_e ratio increased 6 – 9 fold with each 10-fold increase in the selection coefficient in the presence of mutation. However, recombination can break the hitchhiking effect of selection on N_e (Figure 2C). For example, when $r \leq 10^{-3}$, for locus *L* with $U = \mu$, selection with $s = 0.1$ reduced N_e by ~20 fold. In contrast, when $r \geq 0.1$, selection with $s = 0.1$ had little effect on N_e.

Effective population sizes calculated from the inbreeding coefficient (inbreeding N_e) are usually the same as those calculated from the variance in the allele frequency (variance N_e), though exceptions do occur [24-27]. To validate our results, we estimated the effect of selection on N_e by calculating the variance in the frequency of the linked neutral allele from simulations using the same genetic model. Values for the inbreeding N_e obtained from the calculations above were generally consistent with the estimates of the variance N_e derived from these simulations (Figure 2A to 2C). We noted that there was an approximate 3-fold difference in the N_e values between the two methods when $s = 0.01$ (Figure 2B). This is likely due to the fact that the inbreeding N_e was estimated using a strict deterministic model; while the variance N_e was estimated from simulations of $s = 0.01$, where genetic drift plays a bigger role.

A very high mutation rate at the neutral locus *L* (e.g., $U = 1000\mu$) also diminished the reduction in N_e due to selection (Figure 2D). In the absence of mutation, the effect of selection was insensitive to changes in the initial homogeneity at locus *L* (Figure 2E). In the presence of mutation, selection with an initially heterogeneous population at locus *L* caused greater reductions in N_e than selection with

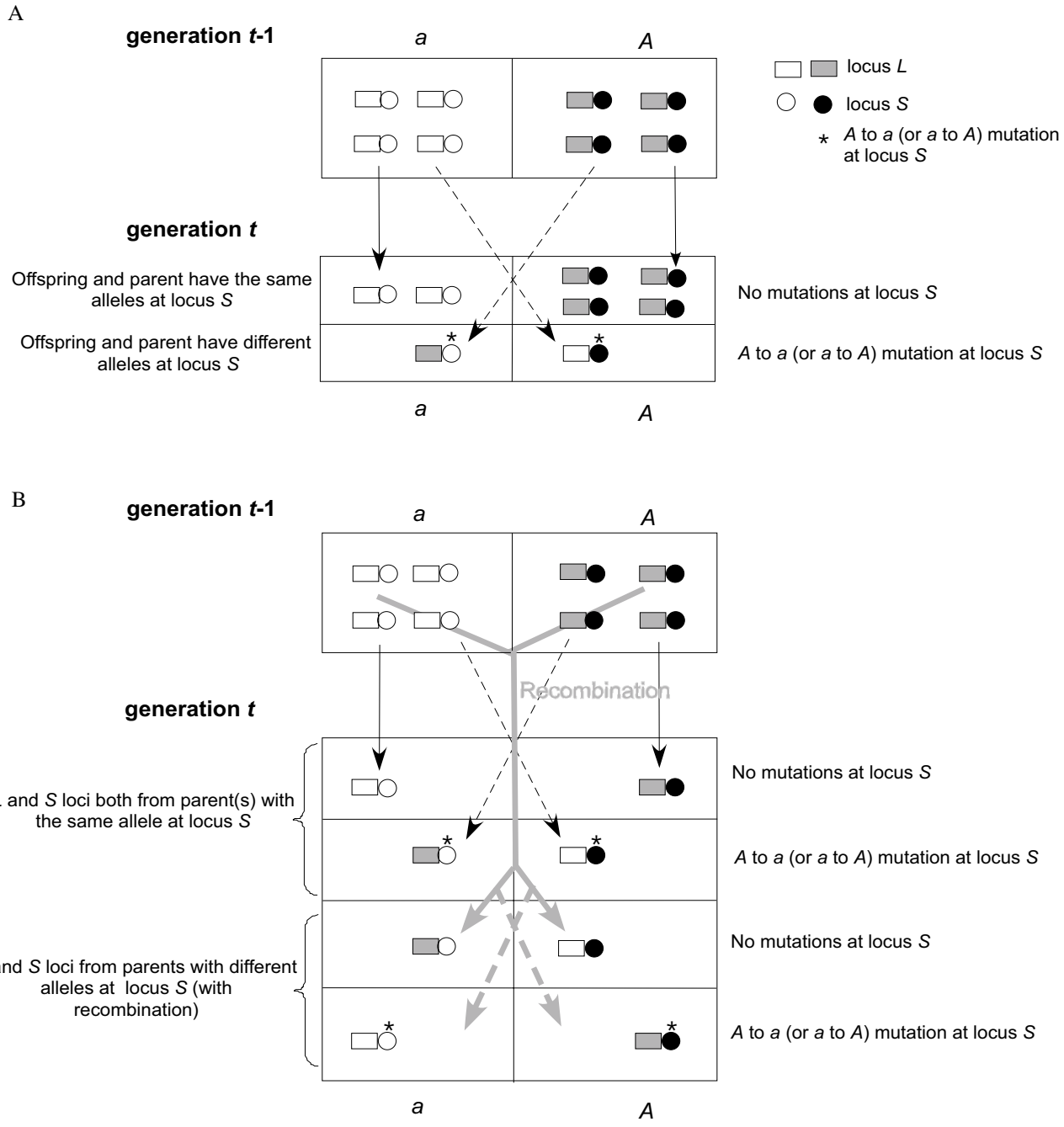

Figure 1

Illustration of the genetic model. A) In the absence of recombination, an offspring with allele *a* can be derived from a parent with allele *a* (without mutation) or a parent with allele *A* (with an *A* to *a* mutation); an offspring with allele *A* can be derived from a parent with allele *A* (without mutation) or a parent with allele *a* (with an *a* to *A* mutation). B) In the presence of recombination, an offspring with allele *a* at locus *S* can be derived from parent(s) in the previous generation by four pathways: 1) Locus *S* from a parent with allele *a* without mutation or recombination, (or with recombination between another parent with allele *a*). 2) Locus *S* from a parent with allele *A* following an *A* to *a* mutation but no recombination (or with recombination between another parent with allele *A*). 3) Locus *S* from a parent with allele *a* without mutation, but with recombination between another parent with allele *A*. 4) Locus *S* from a parent with allele *A* following an *A* to *a* mutation and recombination between another parent with allele *a*. An offspring with allele *A* can be derived from parent(s) in four pathways similar to those described above. For the purpose of illustration, only 8 genomes were presented in generation *t*-1 and *t*.

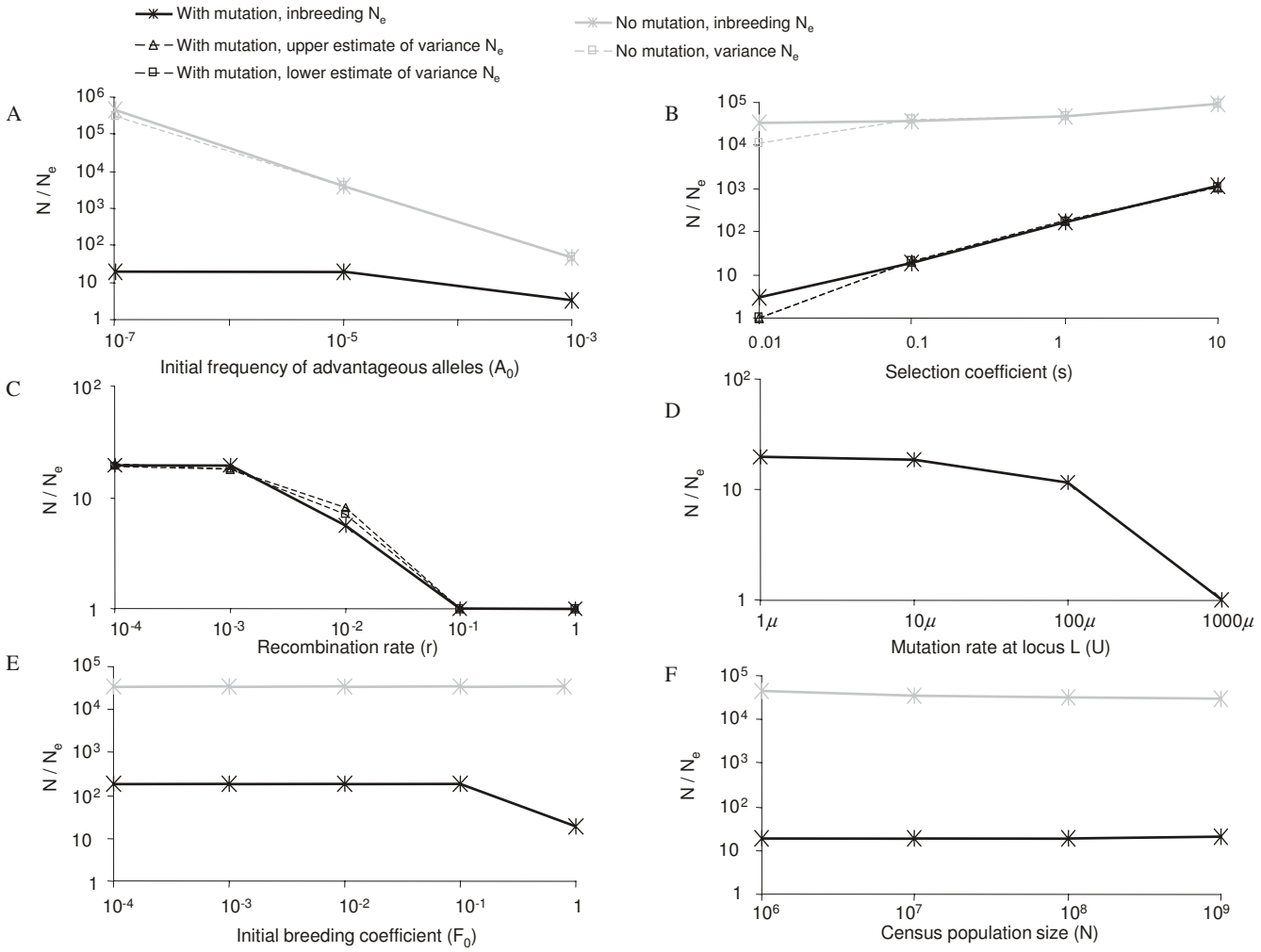

Figure 2

Reduction in N_e following a selective replacement with and without mutations in loci *L* and *S*. A) The effect of different initial frequencies of the advantageous allele. B) The effect of different selection coefficients. C) The effect of different recombination rates. D) The effect of different mutation rates at locus *L*. E) The effect of different initial inbreeding coefficients. F) The effect of different census population sizes. Panels A and C-F all assume $s = 0.1$. Solid lines indicate that the N/N_e ratios are based on the inbreeding coefficient F_t; dashed lines indicate that the N/N_e ratios are based on the variance effective population sizes estimated from our simulations. In the presence of mutation, the dashed lines indicate the N/N_e ratios based on the upper and lower estimates of variance effective populations size. Black lines indicate cases with mutation; grey lines indicate cases without mutation. Unless otherwise specified, the following parameters were used: in the absence of mutation, $\mu = 0$, $v = 0$, $U = 0$, $N = 10^7$, $s = 0.1$, $A_0 = 10^{-6}$, $r = 0$, and $F_0 = F_{AA,0} = F_{aa,0} = F_{Aa,0} = 0.1$; in the presence of mutation, $\mu = 2.5 \times 10^{-5}$, $v = \mu/3$, $U = \mu$, $N = 10^7$, $s = 0.1$, $A_0 = 0$, $r = 0$, and $F_0 = F_{aa,0} = 1$, $F_{AA,0} = F_{Aa,0} = 0$.

an initially homogeneous population. For F_0 less than 0.1, however, further increases in the initial heterogeneity (i.e., making F_0 even lower) did not lead to further reductions in N_e through selection. Interestingly, reductions in N_e/N due to selection were insensitive to changes in the census population size, N (Figure 2F).

Effect of recurrent selection on effective population size

For a homogeneous population under recurrent selection, the inbreeding coefficient of the neutral allele decreased until it reached a quasi-steady state, where it fluctuated in a regular "sawtooth" fashion (Figure 3A). The effect of recurrent selection on N_e was sensitive to selection strength. For example, for a homogeneous population of $N = 10^7$ and $U = \mu$, the decline of F_t over time under recurrent selection with $s = 0.01$ overlapped the neutral curve when $N = 9{,}973{,}000$, while the decline of F_t under recurrent selection with $s = 0.1$ overlapped the neutral curve when $N = 28{,}220$ (Figure 3A). In other words, recurrent selection had little effect on N_e when $s = 0.01$, while recur-

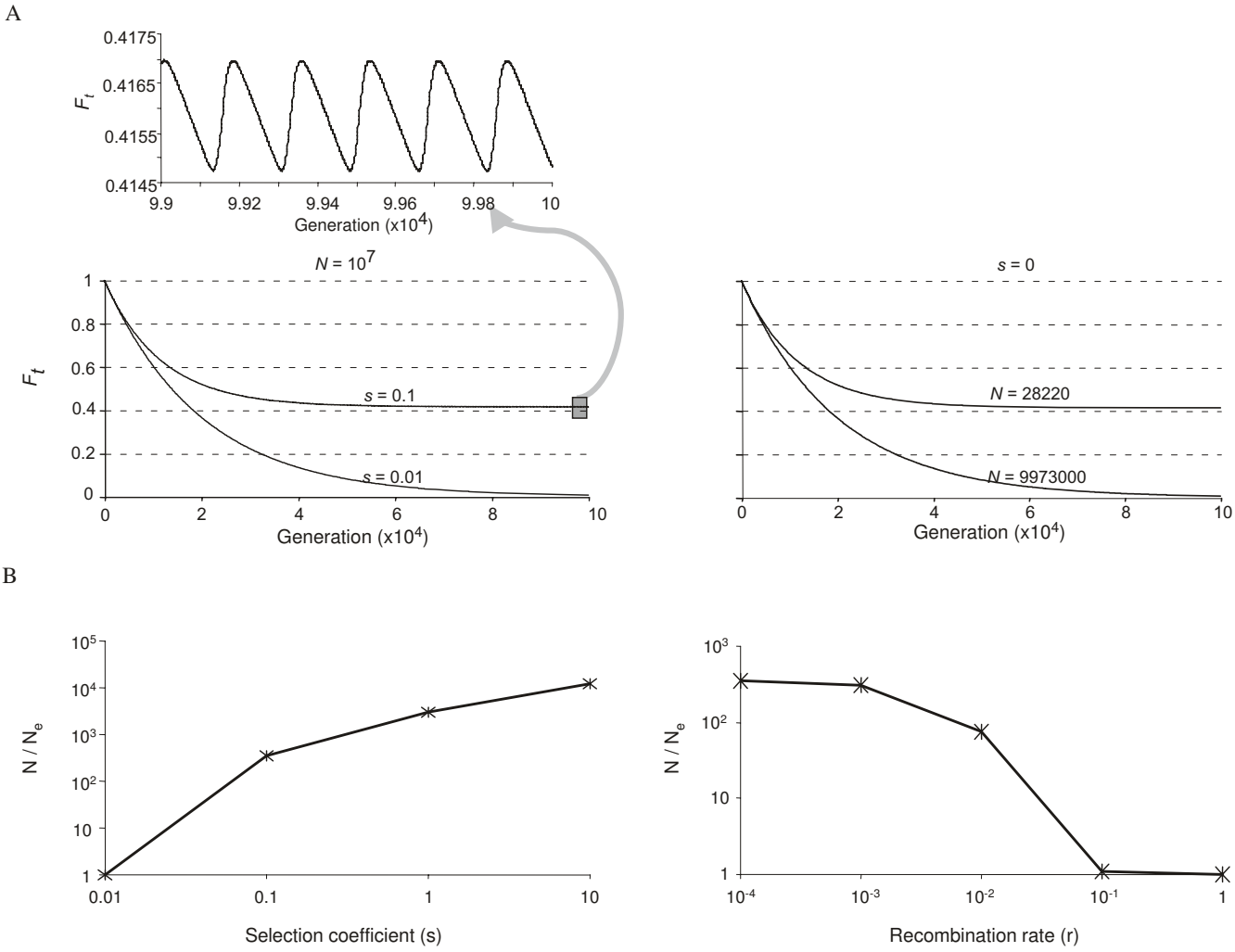

Figure 3

Reduction in N_e due to recurrent selection. A) The changes of F_t over time under selection (left panels) and under neutrality (right panel), in the absence of recombination. B) The effects of different selection coefficients (left panel, $r = 0$) and different recombination rates (right panel, $s = 0.1$). The starting parameters were: $F_0 = F_{aa,0} = 1$, $F_{AA,0} = F_{Aa,0} = 0$, $A_0 = 0$, $U = \mu = 2.5 \times 10^{-5}$, $v = \mu/3$.

rent selection reduced N_e by over 300-fold when $s = 0.1$ (Figure 3A and 3B). This reduction in N_e by recurrent selection could be diminished by high recombination rates (Figure 3B). Although recombination had little impact on the reduction in N_e due to selection under a model with $s = 0.1$ and $r \leq 10^{-3}$, with $r \geq 0.1$, recombination completely broke the hitchhiking effects of selection on N_e with $s = 0.1$.

Conclusion

We examined the combined effects of selection, mutation, and recombination on the effective population size of a neutral locus that is linked to a locus under selection. Consistent with other studies [21-23], we found that selection can increase the inbreeding coefficient and reduce the inbreeding effective population size. Without mutation, this reduction is primarily determined by the initial frequency of the advantageous allele, i.e., the lower the initial frequency, the greater the effect. With mutation, this reduction is mostly determined by the strength of selection, i.e., the stronger the selection, the greater the effect. With moderate recombination rates (e.g., $r \leq 10^{-3}$), recurrent selection can substantially lower N_e, though the hitchhiking effect disappears if the recombination rates are very high (e.g., $r \geq 0.1$).

The effective population size of HIV-1 during chronic infection has been shown to be 100- to 100,000-fold lower than within-host census size. Indeed, CTL responses are a driving force of HIV-1 evolution and these responses

continuously select for escape mutants during chronic infection [28-30]. In a comprehensive study of viral evolution and CTL responses during the first four years of HIV-1 infection in one subject, Liu *et al*. [30,31] found that of the 25 epitopes detected in this subject, 17 were largely replaced by mutants over time. The selection coefficients for the CTL escape mutant(s) of a single epitope ranged from 0.2 to 0.4 during acute infection and from 0.0024 to 0.15 during chronic infection, with an average of 0.03 [30]. Humoral and escape-specific CTL responses impose additional selective pressures not quantified in Liu *et al*. [30,31]. With low to moderate recombination rates, our model shows that recurrent selection with s = 0.1 reduces viral effective population size by approximately 300-fold. Therefore, during HIV-1 infection, selection alone is likely to reduce the viral effective population size to an N_e of ~10^5. This result is close to the estimate of $N_e \sim 5 \times 10^5$ that Rouzine and Coffin [6] obtained from a model that accounts for selection. The small discrepancy may be due to their use of a lower mutation rate (10^{-5} vs. 2.5×10^{-5} in our study) and possible biased sampling of sites with higher underlying mutation rates in their study [5].

With high recombination rates, our model predicts that selection has little effect on N_e. Observations of 3 to 13 cross-over events per virion *in vitro* [17-20] suggest an intrinsic recombination rate of 10^{-4} to 10^{-3} per adjacent site per generation. However, this range is not relevant to our model since these estimates were obtained using heterozygous virions, which may not be abundant *in vivo*. While Jung and colleagues [32] have demonstrated that cells in the spleen are infected with multiple viruses (a pre-requisite for the formation of heterozygous virions), they did not determine how often heterozygous virions are formed. More relevant is data in which SCID-HU mice were infected with a 50:50 mixture of two marked strains [19]. Two-to-three weeks after infection, an average of ~0.01% of infected cells carried a phenotypic marker of recombination (present on half of all recombinants). Conservatively assuming a single generation of recombination, we estimate from equation (11, Appendix) that the probability of recombination between their two markers (which were 408 bp apart) was $r = \sim p_{Aa}/(p_A p_a) = 0.0001/(0.5 \times 0.5) = \sim 4 \times 10^{-4}$ per virion per generation – a value too low to break the hitchhiking effects of selection in our model. However, we recognize these are approximate values obtained from a somewhat artificial system. HIV-1 evolution studies could benefit from additional studies of marked viruses in animal models and clever retrospective analyses of *in vivo* data from humans to determine evolutionarily relevant recombination rates.

Methods

Genetic model

We assume a Wright-Fisher model with a neutral locus L that is linked to a locus under selection, locus S. The selected locus has two alleles, an advantageous allele, A, with a fitness $w = 1 + s$, and a disadvantageous allele, a, with a fitness of 1. Allele A mutates to a at rate μ and allele a mutates to A at rate ν, while neutral mutations at locus L occur at rate U. A description of all the characters, parameters, and variables used in this study is listed in Table 1. For the purposes of calculation, we assume the following parameters are known: the initial frequency of allele A (A_0); the initial frequency of allele a (a_0); and the initial inbreeding coefficient at locus L among all individuals (F_0), among individuals with allele A ($F_{AA,0}$), among individuals with allele a ($F_{aa,0}$), and between individuals with allele A and those with allele a ($F_{Aa,0}$).

Parameters for HIV-1

The average mutation rate of HIV-1 has been estimated to be 2.5×10^{-5} per nucleotide per generation [14], although one recent study estimated a higher mutation rate of ~8.5×10^{-5} per site per generation [15]. Assuming that any nucleotide substitution at a defined nucleotide site shifts locus S from the advantageous to the disadvantageous state, we defined $\mu = 2.5 \times 10^{-5}$ per generation. Assuming that only a particular nucleotide substitution at this site increases fitness, we set $\nu = \mu/3$. Since the census sizes of productively HIV-1 infected cells *in vivo* exceeds 10^7 [7,33], most of the comparisons in this study were with $N = 10^7$. Since the accumulation of advantageous alleles in populations is more stochastic as N decreases, we only examined populations with $N \geq 10^6$.

Effect of selection on effective population size without mutation

Under selection, the inbreeding coefficient of the linked neutral locus will increase faster than expected by random genetic drift until the selected advantageous allele is fixed (A_t = 100%). Because we are using a deterministic model, fixation time is asymptotic. To quantify the effect of selection, we determined the average time for an advantageous allele to approach fixation, $t_{nearlyfixed}$, and the value of F at $t_{nearlyfixed}$. $t_{nearlyfixed}$ can be calculated from $t = \log(\frac{A_t a_0}{a_t A_0}) / \log(w)$, where t is the time just before the favored allele A at locus S becomes fixed; i.e., when $A_t = \frac{N-1}{N}$ and $a_t = \frac{1}{N}$. F was calculated with $\mu = 0$, $\nu = 0$, and $U = 0$. The corresponding N_e is defined here as the population size under neutrality that will increase F from F_0 to $F_{t_{nearlyfixed}}$ between $t = 0$ and $t = t_{nearlyfixed}$. We determined the corresponding N_e under the following condi-

Table 1: Description of characters, parameters and variables.

Characters	Description
S	Locus under selection.
L	Neutral locus linked to locus *S*.
A	Advantageous allele at locus *S*.
a	Disadvantageous allele at locus *S*.
Parameters	
N	Census population size.
s	Selection coefficient.
W	Fitness of the advantageous allele *A*, $w = 1+s$.
μ	Probability that locus *S* mutates from *A* to *a* per virion per generation.
ν	Probability that locus *S* mutates from *a* to *A* per virion per generation.
U	Probability that mutation occurs at locus *L* per virion per generation.
r	Probability of recombination between loci *L* and *S* per virion per generation.
Variables	
N_e	Effective population size.
A_t	Frequency of allele *A* at locus *S* at generation *t*.
a_t	Frequency of allele *a* at locus *S* at generation *t*.
F_t	Probability that two alleles at locus *L* are identical by descent at generation *t* (equivalent to the inbreeding coefficient in classic population genetics).
$\hat{F}$	Inbreeding coefficient at equilibrium.
$F_{AA,\,t}$	*F* of locus *L* between offspring with allele *A*, at generation *t*.
$F_{aa,\,t}$	*F* of locus *L* between offspring with allele *a*, at generation *t*.
$F_{Aa,\,t}$	*F* of locus *L* between offspring with alleles *A* and *a*, at generation *t*.
$p_{A,\,t}$	Probability that an offspring at generation *t* is derived from a parent with allele *A* at generation *t*-1.
$p_{a,\,t}$	Probability that an offspring at generation *t* is derived from a parent with allele *a* at generation *t*-1.
$p_{A\to A,\,t}$	Probability that an offspring at generation *t* is derived from a parent with allele *A*, given that the offspring has allele *A*.
$p_{A\to a,\,t}$	Probability that an offspring at generation *t* is derived from a parent with allele *A*, given that the offspring has allele *a*.
$p_{a\to a,\,t}$	Probability that an offspring at generation *t* is derived from a parent with allele *a*, given that the offspring has allele *a*.
$p_{a\to A,\,t}$	Probability that an offspring at generation *t* is derived from a parent with allele *a*, given that the offspring has allele *A*.
$P_{AA,\,t}$	Probability of an individual at generation *t* having alleles at loci *L* and *S* both being derived from individual(s) with allele *A* at locus *S*.
$P_{Aa,\,t}$	Probability of an individual at generation *t* having locus *L* derived from an individual with allele *A* at locus *S* and locus *S* derived from an individual with allele *a* at locus *S*.
$P_{aa,\,t}$	Probability of an individual at generation *t* having alleles at loci *L* and *S* both being derived from individual(s) with allele *a* at locus *S*.
$P_{aA,\,t}$	Probability of an individual at generation *t* having locus *L* derived from an individual with allele *a* at locus *S* and locus *S* derived from an individual with allele *A* at locus *S*.

tions: $N = 10^6$ to 10^9; $s = 0.01$ to 10; $A_0 = 10^{-7}$ to 10^{-3}; and $F_0 = F_{AA,0} = F_{aa,0} = F_{Aa,0} = 10^{-4}$ to 0.8 (if $F_0 = 1$, F will not change over time without mutation, regardless of selection).

Effect of selection on effective population size with mutation and recombination

The frequency of the *A* allele cannot be maintained at 100% with the occurrence of the back mutation from *A* to *a* at locus *S*. Therefore $t_{nearlyfixed}$ was set to the time that A_t and a_t reached equilibrium, i.e., when $A_t = A_{t+1}$. The corresponding N_e, the population size under neutrality that will increase F from F_0 to $F_{t_{nearlyfixed}}$ between $t = 0$ and $t = t_{nearlyfixed}$, was determined using numerical iteration [Appendix equation (2)]. We determined the corresponding N_e under the following conditions: $N = 10^6$ to 10^9; $s = 0.01$ to 10; $A_0 = 0$ to 10^{-3}; $F_0 = F_{aa,0} = 10^{-4}$ to 1; $F_{AA,0} = F_{Aa,0} = 0$, if $A_0 = 0$ and $F_{AA,0} = F_{Aa,0} = F_0$, if $A_0 > 0$; $\mu = 2.5 \times 10^{-5}$, $\nu = \mu/3$, $U = \mu$ to 1000μ, and $r = 0$ to 1. With these high advantageous mutation rates and large population sizes ($N\nu >> 1$), individuals with allele *a* had mutations to allele *A* in almost every generation, preventing advantageous

allele *A* from being lost from the population due to genetic drift.

Effect of recurrent selection on effective population size

With the fixation of the advantageous allele *A*, the inbreeding coefficient of locus *L* will undergo a nearly neutral change unless new alleles linked to locus *L* become advantageous. To estimate the effect of recurrent selection on N_e, we assumed that all loci under selection are linked to locus *L* in the absence of recombination. We also assumed that each selected locus was under sequential selection, i.e., when the frequency of an advantageous allele reached 99.9% at generation *t*, we assumed that another locus started to undergo selection (calculated by setting $A_t = 0$, $F_{aa,t} = F_t$, $F_{AA,t} = 0$, and $F_{Aa,t} = 0$). For simplicity, we assumed that all of the selected loci have the same mutation rate and selection coefficient. We calculated *F* under recurrent selection under the following conditions: $N = 10^7$, $A_0 = 0$, $F_{AA,0} = F_{Aa,0} = 0$, $F_0 = F_{aa,0} = 1$, $\mu = 2.5 \times 10^{-5}$, $v = \mu/3$, and $U = \mu$; $s = 0.01$ to 10; and $r = 0$ to 1.

Estimate of the effect of selection on variance effective population size by simulation

The change in the average inbreeding coefficient is one of several criteria used to estimate effective population size [24-27]. To validate our results using a different measure of effective population size, we estimated the effect of selection on N_e by calculating the variance in the frequency of the linked neutral allele from simulations using the genetic model described above. The parameters used in these simulations were the same as those used for the calculation for the inbreeding coefficient described above. When simulating selection in the absence of mutation, the simulations were performed under the following conditions: $N = 10^7$; $s = 0.01$ to 10; $A_0 = 10^{-7}$ to 10^{-3}; $F_0 = F_{AA,0} = F_{aa,0} = F_{Aa,0} = 0.1$; $\mu = v = U = 0$; and $r = 0$. When simulating selection in the presence of mutation, the simulations were performed with the following conditions: $N = 10^7$; $s = 0.01$ to 10; $A_0 = 0$; $F_0 = F_{aa,0} = 1$, $F_{AA,0} = F_{Aa,0} = 0$; $\mu = 2.5 \times 10^{-5}$, $v = \mu/3$, and $U = \mu$; $s = 0.01$ to 10; and $r = 0$ to 1. Since the deterministic model assumes an infinite population size, we only examined a large population size of 10^7. For each condition, 100,000 simulations were repeated. We calculated the variance of the allele frequency at the linked neutral locus *L* at the corresponding $t_{nearlyfixed}$. Under neutrality in the absence of mutation, the allele frequency variance can be calculated by $p(1-p)[1-(1-\frac{1}{N})^t]$ [34]. Therefore, the population size under neutrality (N_e) that has the same variance in allele frequency as the population under selection can be determined using numerical iteration. In the presence of mutation, we used simulation to determine the range of the population size under neutrality. These were used to determine the range of allele frequency variances that matched the frequency variance under selection at the corresponding $t_{nearlyfixed}$.

Authors' contributions

YL and JM jointly conceived the study. YL derived the equations, wrote the computer code, performed the computational experiments, and drafted the manuscript. JM advised on the study design, participated in the analysis of the mathematical and computational data, and helped draft the manuscript. Both authors have read and approved the paper.

Appendix

Recurrence equation for **F** *in the absence of selection*

In the absence of mutation or selection, the inbreeding coefficient is

$$F_t = \frac{1}{N} + (1-\frac{1}{N})F_{t-1} = 1-(1-\frac{1}{N})^t(1-F_0) \quad (1)$$

where *t* is time in generations and *N* is the population size [26]. $\frac{1}{N}$ gives the probability that two offspring are derived from same parent in which case the probability of them being identical by descendent is 1. $(1-\frac{1}{N})$ is the probability that two offspring are derived from different parents in which case the probability of them being identical by descendent is F_{t-1}. In the presence of mutation, $F_t = [\frac{1}{N} + (1-\frac{1}{N})F_{t-1}](1-U)^2$ [35]. To obtain F_t in terms of F_0, let $\alpha = \frac{1}{N}\times(1-U)^2$, and $\beta = (1-\frac{1}{N})\times(1-U)^2$. This gives

$$F_1 = \alpha + \beta F_0$$

$$F_2 = \alpha + \beta F_1 = \alpha + \beta \times (\alpha + \beta F_0) = \alpha + \alpha\beta + \beta^2 F_0$$

$$F_3 = \alpha + \beta F_2 = \alpha + \beta \times (\alpha + \alpha\beta + \beta^2 F_0) = \alpha + \alpha\beta + \alpha\beta^2 + \beta^3 F_0$$

$$F_t = \alpha + \alpha\beta + \alpha\beta^2 + \alpha\beta^3 + \ldots + \alpha\beta_{t-1} + \beta^t F_0.$$

The formula, $1 + x + \ldots + x^{n-1} = (1-x^n)/(1-x)$, gives the following:

$$F_t = \alpha(1-\beta^t)/(1-\beta) + \beta^t F_0, \quad (2)$$

As t approaches infinity, F converges to the equilibrium $\hat{F} = \frac{\alpha}{(1-\beta)} \approx \frac{1}{1+2NU}$, as shown previously by Kimura and Crow [35].

Recurrence equations for **F** *in the presence of a selected locus without recombination*

In the presence of selection, the F value at time t is the sum of the probability of two offspring being derived from parents having alleles *AA*, *aa*, or *Aa* at locus *S* multiplied by the probability that the offspring will be identical by descent at locus *L*. In other words:

$$F_t = \left\{ p_{A,t}^2 [\frac{1}{NA_{t-1}} + (1 - \frac{1}{NA_{t-1}}) F_{AA,t-1}] + p_{a,t}^2 [\frac{1}{Na_{t-1}} + (1 - \frac{1}{Na_{t-1}}) F_{aa,t-1}] + 2 p_{A,t} p_{a,t} F_{Aa,t-1} \right\} (1-U)^2. \quad (3)$$

Here, A_{t-1} and a_{t-1} are the frequencies of the advantageous and disadvantageous alleles at locus *S* at generation t-1. $p_{A,t} = \frac{wA_{t-1}}{wA_{t-1}+a_{t-1}}$ and $p_{a,t} = \frac{a_{t-1}}{wA_{t-1}+a_{t-1}}$ give the probabilities that an offspring at generation t is derived from a parent at generation t-1 with allele *A* or *a*, respectively. F_{AA}, F_{Aa}, and F_{aa} give the probabilities that parents with the indicated alleles will be identical by descent at locus *L*. Given that both parents have allele *A* or *a* at locus *S*, the $\frac{1}{NA_{t-1}}$ and $\frac{1}{Na_{t-1}}$ terms respectively give the probabilities that two offspring have the same parent (in which case the probability of being identical by descent at locus *L*, in the absence of mutation is 1). The $1 - \frac{1}{NA_{t-1}}$ and $1 - \frac{1}{Na_{t-1}}$ terms give the probability that the two offspring came from different parents (in which case the probabilities of identity by descent at locus *L*, in the absence of mutation, are $F_{AA,\,t\text{-}1}$ and $F_{aa,\,t\text{-}1}$ respectively). The term $(1\text{-}U)^2$ accounts for the fact that two individuals cannot be identical by descent if there is a mutation at the neutral locus *L*.

If the parameters w, μ, v, U, A_0, a_0, F_0, $F_{AA,0}$, $F_{aa,0}$, and $F_{Aa,0}$ are known, F_1 can be calculated using equation (3). In addition, $F_{AA,1}$, $F_{aa,1}$, and $F_{Aa,1}$ can be calculated using the following equations:

$$F_{AA,t} = \{(p_{A\to A,t})^2 [\frac{1}{NA_{t-1}} + (1 - \frac{1}{NA_{t-1}}) F_{AA,t-1}] + (p_{a\to A,t})^2 [\frac{1}{Na_{t-1}} + (1 - \frac{1}{Na_{t-1}}) F_{aa,t-1}] + 2 p_{A\to A,t} p_{a\to A,t} F_{Aa,t-1}\} (1-U)^2, \quad (4)$$

$$F_{aa,t} = \{(p_{a\to a,t})^2 [\frac{1}{Na_{t-1}} + (1 - \frac{1}{Na_{t-1}}) F_{aa,t-1}] + (p_{A\to a,t})^2 [\frac{1}{NA_{t-1}} + (1 - \frac{1}{NA_{t-1}}) F_{AA,t-1}] + 2 p_{A\to a,t} p_{a\to a,t} F_{Aa,t-1}\} (1-U)^2 \quad (5)$$

and

$$F_{Aa,t} = (p_{A\to A,t} p_{a\to a,t} F_{Aa,t-1} + p_{a\to a,t} p_{a\to A,t} F_{aa,t-1} + p_{A\to a,t} p_{A\to A,t} F_{AA,t-1} + p_{A\to a,t} p_{a\to A,t} F_{Aa,t-1})(1 - U)^2 \quad (6)$$

Where $p_{x\to\gamma,\,t}$ is the probability that a sampled offspring is descended from a parent with allele x given that the offspring has allele γ at locus *S*. The reasoning behind equations (4) – (6) is similar to that for equation (3). For each offspring, $p_{x\to\gamma,\,t}$ can be calculated as the probability of the parent having allele x at locus *S* multiplied by the probability that x mutates to γ (or fails to mutate, if $x = \gamma$), divided by the probability that the offspring is γ. In other words,

$$p_{A\to A,t} = \frac{w\times A_{t-1}}{w\times A_{t-1}+a_{t-1}}(1-\mu)/A_t, \quad p_{a\to A,t} = \frac{a_{t-1}}{w\times A_{t-1}+a_{t-1}} v / A_t$$

and

$$p_{a\to a,t} = \frac{a_{t-1}}{w\times A_{t-1}+a_{t-1}}(1-v)/a_t, \quad p_{A\to a,t} = \frac{w\times A_{t-1}}{w\times A_{t-1}+a_{t-1}} \mu / a_t$$

where A_t and a_t are given by

$$A_t = \frac{w\times A_{t-1}}{w\times A_{t-1}+a_{t-1}}(1-\mu) + \frac{a_{t-1}}{w\times A_{t-1}+a_{t-1}} v \quad (7)$$

$$a_t = \frac{a_{t-1}}{w\times A_{t-1}+a_{t-1}}(1-v) + \frac{wA_{t-1}}{w\times A_{t-1}+a_{t-1}} \mu \quad (8)$$

When A_1, a_1, F_1, $F_{AA,1}$, $F_{aa,1}$, and $F_{Aa,1}$ are available, we can calculate F_2 using equation (3), and $F_{AA,2}$, $F_{aa,2}$, and $F_{Aa,2}$ using equations (4) to (6). Therefore, F_t can be obtained by iteration.

Recurrence equations for **F** *in the presence of a selected locus and recombination*

Assuming that loci *L* and *S* recombine with a probability r per generation, we obtain

$$P_{AA,t} = p_{A,t}(1-r) + p_{A,t} r \frac{Np_{A,t}-1}{N-1} \quad (9)$$

$$P_{aa,t} = p_{a,t}(1-r) + p_{a,t} r \frac{Np_{a,t}-1}{N-1} \quad (10)$$

$$P_{Aa,t} = p_{aA,t} = p_{A,t} r \frac{N p_{a,t}}{N-1} \quad (11)$$

where $p_{xy,\,t}$ is the probability that an individual at generation t has a neutral locus L derived from an individual with allele x at locus S, and a selected locus S derived from an individual with allele y at locus S (x and y can be A or a). This probability is the sum of the probability of no recombination and the probability of recombination between individuals with indicated allele at locus S. The -1's in the ($Np_{x,\,t}$ - 1) and (N - 1) terms above account for the fact that a haploid individual cannot recombine with itself.

Similar to equations (3) – (6), with recombination,

$$F_t = \left\{ \begin{array}{l} (p_{AA,t}^2 + p_{Aa,t}^2 + 2p_{AA,t}p_{Aa,t})[\frac{1}{NA_{t-1}} + (1-\frac{1}{NA_{t-1}})F_{AA,t-1}] \\ +(p_{aA,t}^2 + p_{aa,t}^2 + 2p_{aA,t}p_{aa,t})[\frac{1}{Na_{t-1}} + (1-\frac{1}{Na_{t-1}})F_{aa,t-1}] \\ +2(p_{AA,t}p_{aA,t} + p_{AA,t}p_{aa,t} + p_{aA,t}p_{Aa,t} + p_{aa,t}p_{Aa,t})F_{Aa,t-1} \end{array} \right\}(1-U)^2$$

$$= \left\{ p_{A,t}^2[\frac{1}{NA_{t-1}} + (1-\frac{1}{NA_{t-1}})F_{AA,t-1}] + p_{a,t}^2[\frac{1}{Na_{t-1}} + (1-\frac{1}{Na_{t-1}})F_{aa,t-1}] + 2p_{A,t}p_{a,t}F_{Aa,t-1} \right\}(1-U)^2 \quad (12)$$

$$F_{AA,t} = \{(\frac{p_{AA}(1-\mu)+p_{Aa}v}{A_t})^2[\frac{1}{NA_{t-1}} + (1-\frac{1}{NA_{t-1}})F_{AA,\,t-1}] + (\frac{p_{aA}(1-\mu)+p_{aa}v}{A_t})^2[\frac{1}{Na_{t-1}} + (1-\frac{1}{Na_{t-1}})F_{aa,\,t-1}] + \frac{2[p_{AA}(1-\mu)\times p_{aA}(1-\mu)+p_{AA}(1-\mu)\times p_{aa}v+p_{aA}(1-\mu)\times p_{Aa}v+p_{Aa}v\times p_{aa}v]}{A_t^2} F_{Aa,t-1}\}(1-U)^2 \quad (13)$$

$$F_{aa,t} = \{(\frac{p_{Aa}(1-v)+p_{AA}\mu}{a_t})^2[\frac{1}{NA_{t-1}} + (1-\frac{1}{NA_{t-1}})F_{AA,\,t-1}] + (\frac{p_{aa}(1-v)+p_{aA}\mu}{a_t})^2[\frac{1}{Na_{t-1}} + (1-\frac{1}{Na_{t-1}})F_{aa,\,t-1}] + \frac{2[p_{Aa}(1-v)\times p_{aa}(1-v)+p_{Aa}(1-v)\times p_{aA}\mu+p_{aa}(1-v)\times p_{AA}\mu+p_{AA}\mu\times p_{aA}\mu]}{a_t^2} F_{Aa,t-1}\}(1-U)^2 \quad (14)$$

$$F_{Aa,t} = \{(\frac{p_{AA}(1-\mu)+p_{Aa}v}{A_t})\times(\frac{p_{Aa}(1-v)+p_{AA}\mu}{a_t})[\frac{1}{NA_{t-1}} + (1-\frac{1}{NA_{t-1}})F_{AA,\,t-1}] + (\frac{p_{aA}(1-\mu)+p_{aa}v}{A_t})\times(\frac{p_{aa}(1-v)+p_{aA}\mu}{a_t})[\frac{1}{Na_{t-1}} + (1-\frac{1}{Na_{t-1}})F_{aa,\,t-1}] + \frac{2p_{AA}p_{aA}\mu(1-\mu)+2p_{Aa}p_{aa}v(1-v)+(p_{AA}p_{aa}+p_{aA}p_{Aa})[(1-\mu)(1-v)+\mu v]}{A_t a_t} F_{Aa,\,t-1}\}(1-U)^2 \quad (15)$$

where A_t and a_t are calculated using equations (7) and (8).

Acknowledgements

We thank James I. Mullins for his guidance and critical comments on an early draft, two anonymous reviewers for constructive criticisms, and Reneé Ireton for editing the final version. The authors were supported by grants from the NIH (R03 AI055394, R01 HL072631, P01 AI57005, R01 AI058894, and RO1 AI047734), the University of Washington Center for AIDS Research (NIH grant P30 AI27757), and a gift from the Frank H. Jernigan Charitable Foundation.

References

1. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W: **Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data.** *Genetics* 2002, **161(3):**1307-1320.
2. Leigh-Brown AJ: **Analysis of HIV-1 env gene sequences reveals evidence for a low effective number in the viral population.** *Proc Natl Acad Sci USA* 1997, **94(5):**1862-1865.
3. Achaz G, Palmer S, Kearney M, Maldarelli F, Mellors JW, Coffin JM, Wakeley J: **A Robust Measure of HIV-1 Population Turnover within Chronically Infected Individuals.** *Mol Biol Evol* 2004, **21(10):**1902-1912.
4. Seo T-K, Thorne JL, Hasegawa M, Kishino H: **Estimation of Effective Population Size of HIV-1 Within a Host: A Pseudomaximum-Likelihood Approach.** *Genetics* 2002, **160(4):**1283-1293.
5. Shriner D, Shankarappa R, Jensen MA, Nickle DC, Mittler JE, Margolick JB, Mullins JI: **Influence of random genetic drift on human immunodeficiency virus type 1 env evolution during chronic infection.** *Genetics* 2004, **166(3):**1155-1164.
6. Rouzine IM, Coffin JM: **Linkage disequilibrium test implies a large effective population number for HIV in vivo.** *Proc Natl Acad Sci USA* 1999, **96(19):**10758-10763.
7. Haase AT: **Population biology of HIV-1 infection: viral and CD4+ T cell demographics and dynamics in lymphatic tissues.** *Annu Rev Immunol* 1999, **17:**625-656.
8. Kouyos RD, Althaus CL, Bonhoeffer S: **Stochastic or deterministic: what is the effective population size of HIV-1?** *Trends Microbiol* 2006, **14(12):**507-511.
9. Haddad DN, Birch C, Middleton T, Dwyer DE, Cunningham AL, Saksena NK: **Evidence for late stage compartmentalization of HIV-1 resistance mutations between lymph node and peripheral blood mononuclear cells.** *AIDS* 2000, **14(15):**2273-2281.
10. Fulcher JA, Hwangbo Y, Zioni R, Nickle D, Lin X, Heath L, Mullins JI, Corey L, Zhu T: **Compartmentalization of human immunodeficiency virus type 1 between blood monocytes and CD4+ T cells during infection.** *J Virol* 2004, **78(15):**7883-7893.
11. Wong JK, Ignacio CC, Torriani F, Havler D, Fitch NJS, Richman DD: **In vivo compartmentalization of human immunodeficiency virus: evidence from the examination of *pol* sequences from autopsy tissues.** *J Virol* 1997, **71(3):**2059-2071.
12. Shriner D, Liu Y, Nickle DC, Mullins JI: **Evolution of intrahost HIV-1 genetic diversity during chronic infection.** *Int J Org Evolution* 2006, **60(6):**1165-1176.
13. Barton NH: **Genetic hitchhiking.** *Philos Trans R Soc Lond B Biol Sci* 2000, **355(1403):**1553-1562.
14. Mansky LM: **Forward mutation rate of Human Immunodeficiency Virus Type 1 in a T lymphoid cell line.** *AIDS Res Hum Retroviruses* 1996, **12(4):**307-314.
15. O'Neil PK, Sun G, Yu H, Ron Y, Dougherty JP, Preston BD: **Mutational analysis of HIV-1 long terminal repeats to explore the relative contribution of reverse transcriptase and RNA polymerase II to viral mutagenesis.** *J Biol Chem* 2002, **277(41):**38053-38061.
16. Shriner D, Rodrigo AG, Nickle DC, Mullins JI: **Pervasive genomic recombination of HIV-1 in vivo.** *Genetics* 2004, **167(4):**1573-1583.
17. Zhuang J, Jetzt AE, Sun G, Yu H, Klarmann G, Ron Y, Preston BD, Dougherty JP: **Human Immunodeficiency Virus Type 1 Recombination: Rate, Fidelity, and Putative Hot Spots.** *J Virol* 2002, **76(22):**11273-11282.
18. Jetzt AE, Yu H, Klarmann GJ, Ron Y, Preston BD, Dougherty JP: **High rate of recombination throughout the human immunodeficiency virus type 1 genome.** *J Virol* 2000, **74(3):**1234-1240.
19. Levy DN, Aldrovandi GM, Kutsch O, Shaw GM: **Dynamics of HIV-1 recombination in its natural target cells.** *Proc Natl Acad Sci USA* 2004, **101(12):**4204-4209.
20. Rhodes T, Wargo H, Hu WS: **High rates of human immunodeficiency virus type 1 recombination: near-random segregation of markers one kilobase apart in one round of viral replication.** *J Virol* 2003, **77(20):**11193-11200.
21. Robertson A: **Inbreeding in artificial selection programmes.** *Genet Res* 1961, **2:**189-194.
22. Woolliams JA, Bijma P: **Predicting rates of inbreeding in populations undergoing selection.** *Genetics* 2000, **154(4):**1851-1864.

23. Wray NR, Thompson R: **Prediction of rates of inbreeding in selected populations.** *Genet Res* 1990, **55(1):**41-54.
24. Ewens WJ: **On the concept of the effective population size.** *Theor Popul Biol* 1982, **21:**373-378.
25. Kimura M, Crow JF: **The measurement of effective population numbers.** *Evolution* 1963, **17:**279-288.
26. Wright S: **Evolution in Mendelian Populations.** *Genetics* 1931, **16:**97-159.
27. Hartl DL, Clark AG: **Principles of Population Genetics.** 2nd edition. Sunderland, MA: Sinauer Associates; 1989.
28. Allen TM, Altfeld M, Geer SC, Kalife ET, Moore C, O'Sullivan KM, Desouza I, Feeney ME, Eldridge RL, Maier EL, *et al.*: **Selective escape from CD8+ T-cell responses represents a major driving force of human immunodeficiency virus type 1 (HIV-1) sequence diversity and reveals constraints on HIV-1 evolution.** *J Virol* 2005, **79(21):**13239-13249.
29. Jones NA, Wei X, Flower DR, Wong M, Michor F, Saag MS, Hahn BH, Nowak MA, Shaw GM, Borrow P: **Determinants of human immunodeficiency virus type 1 escape from the primary CD8+ cytotoxic T lymphocyte response.** *J Exp Med* 2004, **200(10):**1243-1256.
30. Liu Y, McNevin J, Cao J, Zhao H, Genowati I, Wong K, McLaughlin S, McSweyn M, Diem K, Stevens C, *et al.*: **Selection on the human immunodeficiency virus type 1 proteome following primary infection.** *J Virol* 2006, **80(19):**9519-9529.
31. Liu Y, McNevin J, Zhao H, Tebit DM, Troyer RM, McSweyn M, Ghosh AK, Shriner D, Arts EJ, McElrath MJ, *et al.*: **Evolution of HIV-1 CTL epitopes: Fitness-Balanced Escape.** *J Virol* 2007, **81(22):**12179-12188.
32. Jung A, Maier R, Vartanian J-P, Bocharov G, Jung V, Fischer U, Meese E, Wain-Hobson S, Meyerhans A: **Multiply infected spleen cells in HIV patients.** *Nature* 2002, **418(6894):**144.
33. Chun T-W, Carruth L, Finzi D, Shen X, DiGiuseppe JA, Taylor H, Hermankova M, Chadwick K, Margolick J, Quinn TC, *et al.*: **Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection.** *Nature* 1997, **387:**183-188.
34. Nei M: **Molecular Evolutionary Genetics.** New York: Columbia University Press; 1987.
35. Kimura M, Crow JF: **The Number of Alleles That Can Be Maintained in a Finite Population.** *Genetics* 1964, **49:**725-738.