Increasing Sequence Correlation Limits the Efficiency of Recombination in a Multisite Evolution Model

S. Gheorghiu-Svirschevski, I. M. Rouzine, and J. M. Coffin
Department of Microbiology and Molecular Biology, Tufts University

The accumulation of preexisting beneficial alleles in a haploid population, under selection and infrequent recombination, and in the absence of new mutation events is studied numerically by means of detailed Monte Carlo simulations. On the one hand, we confirm our previous work, in that the accumulation rate follows modified single-site kinetics, with a timescale set by an effective selection coefficient $s_{\text{eff}}$ as shown in a previous work, and we confirm the qualitative features of the dependence of $s_{\text{eff}}$ on the population size and the recombination rate reported therein. In particular, we confirm the existence of a threshold population size below which evolution stops before the emergence of best-fit individuals. On the other hand, our simulations reveal that the population dynamics is essentially shaped by the steady accumulation of pairwise sequence correlation, causing sequence congruence in excess of what one would expect from a uniformly random distribution of alleles. By sequence congruence, we understand here the opposite of genetic distance, that is, the fraction of monomorphic sites of specified allele type in a pair of genomes (individual sequences). The effective selection coefficient changes more rapidly with the recombination rate and has a higher threshold in this parameter than found in the previous work, which neglected correlation effects. We examine this phenomenon by monitoring the time dependence of sequence correlation based on a set of sequence congruence measures and verify that it is not associated with the development of linkage disequilibrium. We also discuss applications to HIV evolution in infected individuals and potential implications for drug therapy.

Introduction

Genomic recombination is a pervasive biological mechanism. It occurs in living taxa of all levels of complexity, from viruses and prokaryotes to highest order mammals and gymnosperms. Yet its benefits for the survival and propagation of a species are not always clear. Whether recombination is advantageous or not for a certain population seems to depend on a complex set of conditions involving, with equal weight, the nature of intragenome interactions between various beneficial alleles, the structure and size of the population itself, and the type of environment shaping the evolution of the population.

It is frequently argued, for instance, that the exchange of genetic material via recombination (sexual reproduction) expands the range of individual variability that provides the substrate for natural selection (Burt 2000). However, according to few-site models for very large populations in a stationary environment, such an increase in variability arises only under negative linkage disequilibrium (Otto and Lenormand 2002) or negative epistasis (Barton 1995a; Otto and Lenormand 2002). Still because the effects of linkage and epistasis can change with the character of the environment, it was found that recombination may yet increase variability under rapidly changing environmental conditions, regardless of epistasis (Barton 1995a; Burger 1999).

Perhaps the most quoted advantage of recombination in finite populations is that the exchange of alleles between members of a population may promote the accumulation and fixation of beneficial mutations, countering the negative effect of chromosomal linkage on the response to selection (Fisher 1930; Muller 1932). Beneficial mutations survive selection and random drift only if they arise in a sufficiently large number of high-fitness individuals. If the population is not large enough and/or the selection pressure is not high enough, the population will experience an excessive buildup of deleterious mutations known as Muller’s ratchet (Felsenstein 1974), as well as a slow down in the fixation of beneficial mutations that have to compete for representation in the population (clonal interference effect). Recombination has the potential to counteract these effects by bringing together, in the same genome, distinct beneficial mutations that otherwise would likely remain segregated. This old idea (Fisher 1930; Muller 1932) has received quantitative support, both analytically (Crow 1965; Hill and Robertson 1966; Felsenstein 1974; Barton 1995b) and numerically (Hey 1998), typically in models with two or few linked sites. Then again, recombination can just as well dissociate beneficial mutations in highly fit individuals, producing less-fit progeny with diminished chances of survival. In the context of an early deterministic model with multiplicative selection and no linkage disequilibrium, such events were shown to erase entirely the positive effects of recombination on the fixation of beneficial mutations (Maynard-Smith 1968).

The reasons for such divergent conclusions can be traced eventually to various disparities in the assumptions of the different models employed. The enormous complexity of the corresponding in vivo processes calls necessarily for a number of drastic simplifications, starting with the type and size of genome (haploid vs. diploid, 2-allele vs. multiallele, 2 or 3 site vs. multisite or even infinite length), the size of the population (finite vs. infinite, constant vs. variable) and its initial configuration, low or high rate of mutation (when present), type and strength of selection, mechanism and rate of recombination, stochastic or deterministic kinetics, etc. Given the wide range of options and the enduring complexity of the problem, even under strong simplification, it is not surprising that different models lead to different conclusions. The overall impression is, however, that models that take into account the effects of finite population size tend to predict an advantage of recombination in the progressive evolution of organisms (Crow 1965; Hill and Robertson 1966; Felsenstein 1974; Barton 1995b).
One always hopes that approaching the problem with a generous enough set of assumptions will eventually allow a unified understanding of now seemingly antagonistic results.

With this position in mind, we (Rouzine et al. 2003; Rouzine and Coffin 2005) have recently formulated a novel analytical approach to a 2 allele, multisite model of evolution in haploid populations with a simple fitness landscape. Variants of this type of model have proved useful in quantitative studies of the evolution of RNA viruses and bacteria (Tsimring et al. 1996; Aranson et al. 1997; Kessler et al. 1997; Ridgway et al. 1998). The key observation of our approach is that, in a great variety of situations, the distribution of genomes (individual sequences) according to fitness (or the number of less-fit alleles) can be approximated as a quasi-deterministic “solitary wave” with stochastic edges, propagating in fitness space at a slowly changing speed. The method allows the derivation of closed-form expressions for basic kinetic quantities, valid over a broad range of model parameters. Hence, it can provide estimates of the evolution rate for a large variety of theoretical and experimental situations. The framework was first applied to a study of Muller’s ratchet and Fisher–Muller–Hill–Robertson effect under weak selection and mutation (forward, backward, and compensatory), in the absence of recombination (Rouzine et al. 2003). For a wide range of population sizes, it was found that the evolution rate has a logarithmic dependence on the selection coefficient and the population size \( N \). In particular, the model recovered successfully the result of the single-site model in the infinite population limit, the rate of accumulation of beneficial alleles limited by Fisher–Muller–Hill–Robertson effect at intermediate \( N \), a steady state at a critical value of \( N \), and Muller’s ratchet in small populations.

The same approach was employed recently in an analysis of evolution under recombination and weak selection, in the absence of mutations (Rouzine and Coffin 2005), using the approximation that positions of deleterious alleles do not correlate within and between genomes of given fitness. The prediction of a quasi-deterministic solitary wave was found to hold for this model as well. The magnitude of the corresponding accumulation rate followed single-site type kinetics, with the effects of recombination, selection, and random drift combined into an effective selection coefficient \( s_{\text{eff}} \). The effective coefficient \( s_{\text{eff}} \) was smaller than the single-site selection coefficient \( s \), but increased with the recombination rate, confirming that recombination counteracts only partially the Hill–Robertson effect introduced by linkage. Notably, \( s_{\text{eff}} \) and therefore the entire dynamics, was shown to display a well-defined transition with increasing population size \( N \) or recombination rate \( r \). Below a certain critical size \( N_c \sim (N/r)^{1/2} \), the population evolved eventually into a clone and the accumulation of beneficial alleles stopped, corresponding to a null value for \( s_{\text{eff}} \). For population sizes larger than \( N_c \), beneficial alleles accumulated at a steady rate, until such alleles became fixed at all sites. In this case, the effective selection coefficient acquired a logarithmic dependence on \( N \) and approached the single-site selection coefficient \( s \) at large enough \( N \).

The current paper presents detailed numerical tests of the model and conclusions of Rouzine and Coffin (2005) by means of Monte Carlo simulation. We perform a careful numerical analysis of the accumulation kinetics, with specific emphasis on the emergence of intergenome correlation for homologous sites throughout the population. Intergenome correlation is assessed by means of pairwise sequence congruence, defined as the fraction of monomorphic sites of specified allele type in a pair of genomes. Thus, sequence congruence, as a measure of genetic similarity, is the “opposite” of the genetic distance. We examine the interplay between the buildup of sequence congruence and the accumulation of better-fit alleles by monitoring the time dependence of a complementary set of sequence congruence functions and of linkage disequilibrium. Particular attention is also given to the suitability of the single-site form of the accumulation rate expression and to the dependence of the effective selection coefficient on the population size and the recombination rate. As before, we apply our study to the evolution of HIV populations and the probable role of recombination in the emergence of drug resistance.

Model and Methods

Model

We model the viral population as a collection of \( N \) haploid genomes, represented by sets of \( L \) sites (bases or base pairs) each and subject to discrete time Wright–Fisher evolution with recombination, random drift, and selection for better-fit genotypes. For simplicity, each site is assumed to display only one of 2 alleles: better fit or less fit. We further disregard any epistasis, as well as all mutation events. Deleterious mutation events are neglected under the assumption that the accumulation rate of better-fit sites, in the parameter range of interest, is much larger than the mutation rate. Beneficial mutation events are also neglected because all beneficial alleles are assumed to preexist. This simplification allows us to consider only significant sites that present at least one copy of a better-fit allele in the initial population. The appearance of new significant sites at later times can be ignored, and the effective genome length \( L \) is fixed accordingly for the duration of the evolution. The population size \( N \) is also assumed to be constant in time.

The initial population is subjected to a discretized evolution with a time step of one generation. Each generation, a (small) number of randomly sampled pairs of genomes undergoes recombination, and the entire population is subsequently replaced by sampled progeny. In the sampling process, individual genomes are substituted by their direct descendants in random numbers according to a Poisson distribution, under the constraint of a constant population size. The average number of progeny for any given genome is \( \lambda \) per generation. We model the viral population as a collection of \( N \) haploid genomes, represented by sets of \( L \) sites (bases or base pairs) each and subject to discrete time Wright–Fisher evolution with recombination, random drift, and selection for better-fit genotypes. For simplicity, each site is assumed to display only one of 2 alleles: better fit or less fit. We further disregard any epistasis, as well as all mutation events. Deleterious mutation events are neglected under the assumption that the accumulation rate of better-fit sites, in the parameter range of interest, is much larger than the mutation rate. Beneficial mutation events are also neglected because all beneficial alleles are assumed to preexist. This simplification allows us to consider only significant sites that present at least one copy of a better-fit allele in the initial population. The appearance of new significant sites at later times can be ignored, and the effective genome length \( L \) is fixed accordingly for the duration of the evolution. The population size \( N \) is also assumed to be constant in time.

The average number of progeny for any given genome is proportional to its fitness, which is determined on the assumption that each expressed nonresistant allele diminishes the chances of survival by a factor \( e^{-s} \), slightly less than unit \( (s \ll 1) \). In other words, log fitness is proportional to the number \( k \) of less-fit sites, by the selection coefficient \( s \) \( (\log \text{fitness} = -sk, \text{no epistasis}) \).

The model recombination process assumes that recombiant genome pairs are selected randomly and sequentially with a probability \( 1/N_p \), such that each genome has an equal chance of undergoing recombination, at a rate \( r = N/N_p \). Each generation, the randomly sampled genome pairs undergo a preset number of sequence crossovers to
produce recombinant pairs of descendant strands. Both descendants of a recombinant pair are retained in the evolving population (even though in a real population only one descendant is produced, the difference hardly matters, as only unusually highly fit recombinants influence evolution). We note that because of this latter assumption, our parameter \( r \) should not be thought of as a probability, because it may well happen that \( r > 1 \). In such a case, there is a nonzero probability that some descendants of genomes already involved in a recombinant event become themselves involved in recombination within the same generation. In the case of HIV, the rate \( r \) is determined by the frequency of coinfection of a target cell by 2 (or more) viral particles. More precisely, \( r \) represents the average number of coinfection events experienced by a target cell during one generation. Assuming that each such coinfection event results in the recombination of the 2 viral strands, the average number of recombinant encounters per generation in the entire population amounts to \( \sim rN/2 \).

Model Parameters for HIV Populations In Vivo

The model and the model parameters are tailored to fit our particular interest in HIV evolution. The details depend on the timescale considered. For long observation periods, \( 10^2 \) to \( 10^3 \) generations, the number of polymorphic sites is on the order \( L \sim 100 \). For short-term evolution, the effective genome length \( L \) (number of drug-sensitive polymorphic sites) may be much smaller. The effective size of an HIV population, \( N \), is set by the number of infected cells on the productive pathway. According to existing estimates based on models that include selection (Chun et al. 1997; Rouzine and Coffin 1999; Frost et al. 2000), in an average patient, this number is \( N \sim 10^6 \) or larger. However, the time required for our Monte Carlo simulations increases at least quadratically with the population size, and simulations become rather inefficient at \( N > 10^3 \). We compromise on somewhat smaller population sizes, \( 10^5 \) or fewer. A similar predicament constrains our choice of allele composition. We consider cases when only a small fraction \( f_0 \) of the alleles presented by significant sites at the onset of evolution are better fit. This initial condition corresponds to 2 experiments: 1) reversion of deleterious alleles in untreated patients soon after transmission due to the transmission bottleneck and 2) amplification of drug-resistant alleles after start of drug therapy. But a very low initial fraction of better-fit alleles requires higher \( N \) to avoid their loss, \( Nf_0 \gg 1 \), and leads again to inefficiently long simulation times. We prefer to balance against lengthy simulations by considering initial compositions with \( f_0 \sim 10^{-2} \) randomly distributed drug-resistant alleles. In the absence of initial allele loss, simulation results are weakly sensitive to \( f_0 \).

Variation of the selection coefficient \( s \) among different sites is ignored, replaced by a representative value determined by the timescale of the experiment under consideration \( (t \sim 1/s) \). Indeed, given a timescale, beneficial alleles at sites with much larger \( s \) will be fixed rapidly and similar alleles at sites with much smaller \( s \) will not have enough time to accumulate. Short-term evolution, spanning \( 10-100 \) generations, suggests \( s \sim 0.1 \), whereas longer timescales, stretching over hundreds and thousands of generations, imply \( s \sim 10^{-2} \) (Rouzine and Coffin 1999). Much smaller values of \( s \) may be dismissed because the corresponding timescales exceed the average duration of HIV infections. Larger values \( s \sim 1 \), corresponding to short durations of the order of several generations, are also irrelevant because such sites can be regarded as conserved on longer timescales. Our simulations use typical values of \( s = 0.1-0.01 \).

In each round of infection, 2 RNA templates carried by a virus particle are reverse transcribed into a new provirus. Recombination occurs when the reverse transcriptase switches between the 2 available templates and incorporates fragments of genetic material from both parental strands into the newly transcribed DNA genomes. It is assumed that a relatively large number, \( \sim 10 \), of such switches (crossovers) occurs before the reverse transcription process is completed (Dang et al. 2004; Levy et al. 2004), with the result that the progeny provirus carries a well-mixed blend of sequences from the parental RNA strands. Two different virus genomes can recombine only if they coinfect the same cell. If the density of newly infected cells is low in the tissue, as in persistent HIV infection, the rate \( r \) of recombination for any one genome becomes proportional to the number \( N \) of infected cells, \( r = N/N_0 \). The factor \( 1/N_0 \) represents the probability that any randomly selected pair of genomes succeeds in coinfection; alternatively, \( N_0 \) is the extrapolated population size for which the recombination rate becomes 1. It depends on the total number of cells permissive for virus replication and their distribution in tissue. Whereas an estimate, \( r \sim 1 \), has been obtained in some untreated patients based on the sampling of double HIV DNA positive cells (Jung et al. 2002), conclusive experimental measurements of the rate \( r \) in productively infected (RNA positive) cells for given \( N \) have yet to be performed. Unfortunately, this leaves a good deal of uncertainty concerning the relevant values for both \( N \) and \( r \) in real patients. For instance, the above estimate of \( r \) may be too high if most coinfected virions originate in neighboring cells and therefore are genetically uniform. Similarly, the estimate for \( N \), which is based on a 2-site model, may have to be adjusted both to account for a higher frequency of observed crossovers and for the multisite nature of evolution. In the following, we use values on the order \( N_0 \sim 10^3 \) to \( 10^5 \), which may be lower than the typical value in an untreated patient but offer a reasonable compromise in terms of simulation time. We note that \( N_0 \) can be varied in vitro by changing the culture conditions, specifically, the number, density, and type of permissive cells. The model is equally applicable for other partially sexual populations, such as yeast, in which the recombination rate \( r \) is regulated independently of \( N \) by external factors.
time-dependent frequency of genomes with \(k\) less-fit alleles, \(f(k, t)\). It has been found previously (Rouzine et al. 2003; Rouzine and Coffin 2005) that, under the assumptions discussed above, the main part of the distribution \(f(k, t)\) follows a quasi-deterministic kinetics. By contrast, the edges of the distribution (beyond which all fitness classes are empty), especially the leading edge toward small \(k\), advance at a rate determined by the stochastic formation of best-fit sequences through recombination. Away from the edges \(f(k, t)\) is expected to follow kinetics of the form

\[
\frac{df(k, t)}{dt} = (e^{-s(k-t)} - 1)f(k, t) + R(k, t), \tag{1}
\]

where the first term on the right hand side describes the selection process, whereas the generic term \(R(k, t)\) subsumes the effects of recombination. The quantity \(e^{-sk} = \sum_k e^{-sk}f(k)\) accounts for the normalization of the selection contribution to constant population size. If the profile of \(f(k, t)\) is always narrow enough that \(s(k-1) < 1\), then the normalization factor becomes \(e^{sk} = \sum_k f(k, t)\), where \(k\) is the average number of less-fit alleles, and the kinetic equation (1) acquires the simpler form

\[
\frac{df(k, t)}{dt} = -s(k-\langle k \rangle)f(k, t) + R(k, t). \tag{2}
\]

In this limit, we retrieve immediately the well-known Fisher’s law for the evolution of \(\langle k \rangle\):

\[
\frac{dk(t)}{dt} = -s\text{Var}[k], \tag{3}
\]

with \(\text{Var}[k] = \langle(k - \langle k \rangle)^2\rangle\) being the variance of the number \(k\) of deleterious alleles (Haigh 1978). The recombination term \(R(k, t)\) does not appear in equation (3) because the recombination process does not change the number of alleles of one type or another, as given by \(\int dk k R(k, t) = 0\), and \(R\) does not contribute directly to changes in \(\langle k \rangle\). Its contribution is concealed in the explicit form of the variance \(\text{Var}[k]\).

At infinite \(N\) or small \(N_0\), the expression of \(\text{Var}[k]\) can be further inferred without direct knowledge of the recombination term \(R\), recalling that in this case the initial random (binomial) distribution of mutations \(\langle k(t=0)\rangle = L - 1\) remains random (binomial) throughout the evolution. Indeed, in either case, the average number of recombination events per generation, \(rN/2 = N^2/2N_0\), becomes infinitely large, and such strong recombination maintains a perfectly random (binomial) redistribution of alleles throughout the population. As in the single-site case, the corresponding variance is given by

\[
\text{Var}[k] = \langle k \rangle(1 - \langle k \rangle/L). \tag{4}
\]

Substitution of equation (4) into equation (3) yields a standard deterministic evolution equation for the single-site allelic frequency \(\langle k \rangle/L\).

At finite \(N\) and large \(N_0\), the binomial distribution of alleles, (eq. 4), and therefore the deterministic expression for the evolution rate, equation (3), does not apply directly for 3 reasons as follows.

(i) Because better-fit alleles at different sites are selected concurrently and because new sequences are produced at a limited rate, the number \(k\) of less-fit sites correlates between different genomes. Recombination produces best-fit individual genomes that spread gradually over the population, then the process is repeated. As a result, at a given \(\langle k \rangle\), the distribution of \(k\) among genomes is more narrow than given by equation (4), and the evolution rate, correspondingly, is smaller than that given by equation (3). In order to determine the correct variance \(\text{Var}[k]\), one has to specify the expression for the recombination term \(R(k, t)\) in equation (2). In the approximation that the positions of individual sites are completely random within a genome and, unlike their total number, do not correlate between different genomes, \(R(k, t)\) can be expressed in terms of \(f(k, t)\), as a quadratic functional (Rouzine and Coffin 2005). Matching the rate of generation of new best-fit genomes (i.e., those with smallest \(k\)) to the evolution rate of the deterministic part of fitness distribution, we (Rouzine and Coffin 2005) obtained

\[
\text{Var}[k] = \left(\frac{s_{eff}}{s}\right)\langle k \rangle(1 - \langle k \rangle/L), \tag{5a}
\]

where \(s_{eff}\) is given by

\[
s_{eff} = \frac{\ln(Nr)}{\ln[(Ns)^2\langle k \rangle(1 - \langle k \rangle/L)/\langle Nr \rangle)]}, \tag{5b}
\]

or, taking into account that \(r = N/N_0\),

\[
s_{eff} = \frac{\ln(N/N_0)}{\ln[s\sqrt{Ns}\langle k \rangle(1 - \langle k \rangle/L)/\langle Nr \rangle)]}. \tag{5c}
\]

Thus, the binomial dependence of \(\text{Var}[k]\) on \(\langle k \rangle\), equation (4), is mostly preserved, but \(\text{Var}[k]\) is decreased by a factor \(s_{eff}/s\) that becomes one at \(N = \infty\) or \(r \sim s\sqrt{Ns}\langle k \rangle(1 - \langle k \rangle/L)\), if \(r\) and \(N\) are independent, and at \(N = N_0s\sqrt{Ns}\langle k \rangle(1 - \langle k \rangle/L)\), if \(r\) and \(N\) are proportional. Accordingly, the deterministic evolution equation (3) formally holds, but with a smaller effective selection coefficient \(s_{eff}\) that absorbs the effects of finite population size, linkage (Hill–Robertson effect), and recombination.

(ii) The approximation (Rouzine and Coffin 2005) that the positions of beneficial alleles, for a given number per sequence, do not correlate between genomes may be inaccurate. Even if the initial distribution of alleles is completely random, correlations between genomes may develop later because pairs of recombining genomes share sequence segments inherited from common ancestors. Suppose, for example, that 2 genomes are produced by recombination between 2 pairs of parental genomes. If one parent sequence is common for both recombinants, approximately a quarter of their sequence will be identical. The frequency of common ancestral sequences is limited by the combined effect of selection and random sampling. At this time, we are not able to provide good theoretical estimates for the magnitude of this effect. However, the direct numerical simulations presented below will show that the effective selection coefficient
approximation provides very good estimates for the time
dependence of \( \langle k \rangle \) and the variance \( \text{Var}[k] \), but equation
(5b) for \( s_{\text{fix}} \) has to be modified substantially, by a factor
that is a nontrivial functional of the model parameters \( N, L, r, s \).

(iii) equation (5a) for the variance \( \text{Var}[k] \) has to be adjusted,
in addition, for the fixation and loss of better-fit alleles
because monomorphic sites do not contribute to fluctuations
of \( k \). If \( k_{\text{lost}}(t) \) and \( k_{\text{fix}}(t) \) denote, respectively, the
number of lost and fixed better-fit sites at time \( t \), these
modifications bring equation (5a) for \( \text{Var}[k] \) to the form

\[
\text{Var}[k] = \left( \frac{s_{\text{fix}}}{s} \right) \langle k \rangle - k_{\text{fix}}(t) \left( 1 - \frac{\langle k \rangle - k_{\text{fix}}(t)}{L - k_{\text{lost}}(t) - k_{\text{fix}}(t)} \right). \tag{6}
\]

Numerical Implementation

The Monte Carlo simulations were performed with our
newly developed microPopulation software package
(Gheorghiu-Svirschevski S, Rouzine I, unpublished data).
The corresponding algorithm of the Wright–Fisher kinetics
is rather simple. The model population is formally represen-
ted by an array of \( N \) strings with \( L \) binary variables each.
The initial array is constructed site by site according to a
binomial distribution with an average of \( f_0 \cdot L \) beneficial al-
leles per genome. Each Monte Carlo generation (iteration)
carries the population array through one session of recom-
bination events and one session of selection. During the
recombination step, parent genome pairs, selected sequen-
tially and randomly with a probability \( 1/N_0 \), undergo a prese-
number of \( M = 10 \) crossovers each, at randomly selected
positions along the genome length. No recombination hot-
spots are used in our simulations. (A random number of
crossover events would be a better assumption. However,
because the results are not sensitive to the value chosen for
\( M \) either). The resulting population is then
submitted to a selection algorithm based on the well-known
“broken stick” algorithm. The unit interval \([0, 1]\) is divided
into \( N \) segments, each segment corresponding to one ge-
nome in the population. The length of each segment is
set proportional to the fitness of the matching genome. Then
the \( N \) genomes surviving into the next generation are se-
lected by \( N \) random drawings of a random variable uni-
formly distributed on \([0, 1]\), according to the particular
segment on which each drawing places the random variable.

Calculated Quantities

Our numerical calculations have targeted a 3-fold task.
First, we performed a detailed numerical check on the ac-
tual validity of the inferred kinetics described by equation
(6). The distribution \( f(k, t) \) of genomes with given number \( k \)
of less-fit alleles was monitored throughout the run, along
with the corresponding average \( \langle k \rangle \), its time derivative
\( d\langle k \rangle/dt \), and the variance \( \text{Var}[k] \). For the same purpose,
we also monitored the time dependence of the number of lost
and fixed better-fit alleles. Second, we paid particular atten-
tion to similarity in pairs of sequences. To give quantitative
expression to this phenomenon, we defined 3 complemen-
tary measures of the degree of sequence congruence (the
opposite of the genetic distance) in the following way.
For any given pair of genomes \( i \) and \( j \), we denote \( n_{\text{bf}} \) and \( n_{\text{ff}} \)
the number of monomorphic less-fit and better-fit sites, respec-
tively, and denote \( n_{\text{poly}} \) the number of polymorphic sites.
Observe that \( n_{\text{bf}} + n_{\text{ff}} + n_{\text{poly}} = L \). For each type of
allele, we define a pair congruence measure \( C_{\text{bf}}(i, j) \) as
the ratio of monomorphic sites for that type of allele to the
total number of sites expressing the type in at least one of
the genomes. That is,

\[
C_{\text{bf}}(i, j) = \frac{n_{\text{bf}}}{n_{\text{bf}} + n_{\text{ff}} + n_{\text{poly}}} = \frac{n_{\text{bf}}}{L - n_{\text{ff}}} = \frac{n_{\text{bf}}}{n_{\text{ff}} + n_{\text{poly}}} = \frac{n_{\text{bf}}}{L - n_{\text{ff}}}. \tag{7}
\]

Then each degree of sequence congruence is averaged over
all possible pairs in the population,

\[
C_{\text{bf}} = \langle C_{\text{bf}}(i, j) \rangle, \quad C_{\text{ff}} = \langle C_{\text{ff}}(i, j) \rangle. \tag{8}
\]

We also define a total sequence congruence \( C_{\text{tot}} \) by taking
into account all monomorphic sites, irrespective of the type
of allele expressed. In this case, \( C_{\text{tot}}(i, j) = n_{\text{bf}} + n_{\text{ff}} \)
and

\[
C_{\text{tot}} = \langle C_{\text{tot}}(i, j) \rangle, \tag{9}
\]

where the average is again taken over all possible pairs of
genomes. Note that all measures defined in equations
(8–9) are positive and normalized to 1, \( 0 \leq C_{\text{bf}}, \ C_{\text{tot}} \leq 1 \). Value \( 1 - C_{\text{tot}} \) is the average pairwise genetic
distance.

In our third numerical task, we checked the time de-
dependence of linkage disequilibrium throughout the model
dynamics. Linkage disequilibrium was monitored via the
2-site Lewontin measure \( D'(\alpha, \beta) \) (Lewontin 1964) aver-
eged at every given time over all pairs of sites separated
by a prescribed mutual distance \( d, D' = \langle D'(\alpha, \alpha + d) \rangle \).
Here \( \alpha \) and \( \beta = \alpha + d, 1 \leq \alpha, \beta \leq L \), label 2 arbitrary sites of the
viral genome separated by \( d \) intermediate positions. The
Lewontin disequilibrium is calculated as follows. If
\( f_{\alpha 1} (\alpha, \beta) \) denotes the probability that both site \( \alpha \) and site
\( \beta \) contain less-fit alleles, \( f_{\alpha 1} (\alpha) \) denotes the probability of
a less-fit allele at site \( \alpha \) alone, and \( D(\alpha, \beta) = f_{\alpha 1}(\alpha, \beta) -
f_{\alpha 1}(\alpha)f_{1}(\beta) \), then

---

**Fig. 1.** Monte Carlo simulations of the evolution of genomes: (a) Time dependence of the average distribution of genomes with \( k \) less-fit mutations, \( f(k, t) \), for the listed parameters. The recombination rate per genome is \( r = N/N_0 = 0.5 \), where the characteristic population size is \( N_0 = 10^7 \). The average was computed over 10 runs, with identical initial conditions, \( \langle k \rangle = L - 1 \), modeling the initial state of an HIV population in untreated low-viremia patients or at the beginning of antiretroviral therapy. Alternating colors label successive times in generations. (b) Time dependence of the average number of less-fit alleles \( \langle k \rangle \), the normalized reversion rate \( s^\gamma d\langle k \rangle/dt \), the variance of \( k \), \( \text{Var}[k] \), and the time dependence of the average numbers of lost and fixed less-fit alleles for the parameter set shown in (a). Color coding is shown. (c) Same as in (b), but for a smaller population size. Parameters are shown. The recombination rate per genome is \( r = N/N_0 = 0.3 \). (d) Variation with the population size \( N \) of the average mutation number \( \langle k_{\text{final}} \rangle \) of final clones and of the width \( \text{std}(\langle k_{\text{final}} \rangle) \) of the corresponding distribution. \( L, s, \) and \( N_0 \) are shown.
Results
Kinetics of Better-Fit Allele Accumulation

The typical observed evolution of the distribution $f(k, t)$ for a sufficiently large population size $N$ and random initial distribution of alleles is shown in figure 1a (average over 10 runs with an identical initial population). Parameter values are defined in the figure. Figure 1b shows the corresponding time dependence (averaged over 10 runs) of $\langle k \rangle$, $s^{-1}|d\langle k \rangle/dt|$ and $\text{Var}[k]$, as well as the time-dependent numbers of lost and fixed better-fit alleles, $k_{\text{lost}}$ and $k_{\text{fix}}$. The 3 allele-based measures of sequence congruence obtained numerically using equations (7–9) are shown in figure 2a, plotted against the current $\langle k \rangle$. Note that $\langle k \rangle(t)$ is a strictly decreasing function of time, and thus the arrow of time in figure 2a runs from larger $\langle k \rangle$ to lower $\langle k \rangle$. Estimates of the variance $\text{Var}[k]$ from equation (6) are compared with numerical values from simulations in figure 2b. The estimates were calculated using $s_{\text{eff}}$ as a fitting parameter and the numerical values of $k_{\text{lost}}$ and $k_{\text{fix}}$. Also shown in figure 2b are estimates of the variance $\text{Var}[k]$ in terms of the total sequence congruence $C_{\text{tot}}(t)$ (see eq. 11) below.

It can be easily observed from figure 1 that the presence of recombination, under selection and random drift fosters a steady accumulation of better-fit alleles from one generation to the next. One can distinguish 3 significant dynamic phases:

Phase 1. The quasi-singular distribution corresponding to the largely less-fit initial population $\langle k \rangle_{t=0}\approx 0.99L$ disperses into a bell-shaped profile that slowly shifts toward lower values of $k$ (see fig. 1a and b). This distribution corresponds to a rapid initial accumulation of better-fit alleles and thus to a substantial diversification of the population.
By the end of this transient period, the distribution \( f(k) \) develops a relatively large number of individual \( k \)-groups, most comprising a sizeable number of genomes. Phase 2. The bell-shaped distribution continues to shift toward lower numbers of \( k \), this time at a considerably accelerated but slowly varying rate (fig. 1a and b). Notably, the distribution profile varies only slightly during this phase (fig. 1a, 40 \( \leq t \leq 100 \)); that is, as pointed out in our earlier work (Rouzine et al. 2003; Rouzine and Coffin 2005), the propagation of \( f(k, t) \) can be approximated as a solitary wave.

Phase 3. The remaining less-fit alleles continue to be eliminated until (almost) completely purged. In other words, the better-fit alleles undergo gradual fixation, whereas the bell-shaped profile of \( f(k) \) shrinks until it eventually reduces to a single stationary group at \( k_{\text{final}}=0 \).

At lower population sizes, this general pattern is altered by the loss of better-fit alleles during Phase 2 (fig. 1c). In this case, the ongoing accumulation of better-fit alleles is limited by the diminished number of polymorphic sites. Beyond the solitary wave regime, the loss of better-fit alleles declines rapidly until it eventually halts, whereas their fixation also begins at random available sites. As in the case of larger \( N \), the latter process continues until complete elimination of nonfixed less-fit alleles, with the result that the entire population is reduced finally to a clone, with \( k_{\text{final}} \) 0 randomly positioned less-fit sites (fig. 1a). The number \( k_{\text{final}} \) of less-fit sites surviving in any particular final clone samples a bell-shaped distribution centered on an average position \( \langle k_{\text{final}} \rangle \) (results not shown). Figure 1d shows that, for fixed recombination and selection rates, \( \langle k_{\text{final}} \rangle \) depends on the population size \( N \). It decreases toward \( k=0 \) with increasing \( N \), whereas the width of its corresponding distribution shrinks, such that in sufficiently large populations the final clone displays invariably the best-fit genome that can evolve \( (k_{\text{final}}=0) \).

As was checked in all simulations that we have run, the kinetics of allele accumulation proves to be in excellent agreement with the general prescription of Fisher’s law, equation (3), that the speed of evolution is proportional to the width of the distribution \( f(k, t) \) (see 2 examples in fig. 1b and c). Figure 2b shows that we also find very good agreement with the modified one-site kinetics of equation (6), based on the effective selection coefficient approach. The dependence of \( s_{\text{eff}} \) on \( N \) and other parameters is further discussed in the next section.

Sequence Congruence

The emergence of congruent sequence segments throughout the population was followed by means of the 3 measures defined in equations (7–9). A typical output is shown in figure 2a, plotted against the current average number of less-fit sites \( \langle k \rangle \).

The overall sequence congruence \( C_{\text{int}} \) and the congruence of less-fit sites \( C_{\text{lf}} \) both display a nonmonotonous dependence, with a rather shallow minimum between maximal unit values at the beginning and at the end of the evolution (fig. 2a). Unit congruence corresponds to (almost) uniform final and initial populations. During most of the evolution, the congruence of less-fit sites \( C_{\text{lf}} \) decreases due to the constant accumulation of better-fit alleles. The final increase in \( C_{\text{lf}} \) is related to the shrinking of the distribution \( f(k, t) \) (fig. 1a) and the emergence of the final clone. The increase of total congruence during the second half of evolution witnesses the increasing congruence of better-fit sites.

The congruence of better-fit sites \( C_{\text{bf}} \) increases monotonously with \( \langle k \rangle \) throughout the evolution, from an initial null value for \( \langle k \rangle \approx 99 \) \( (f_0 = 0.01, L = 100) \) at \( t = 0 \) to the unit value marking the establishment of the final clonal population. Interestingly, in the parameter range we tested, for example, \( N_0 = 10^4 \), the time dependence of \( C_{\text{bf}}(t) \) is approximately linear throughout the evolution. Although we do not have a theoretical understanding of this feature, it is worth pointing out that the approximately linear dependence of \( C_{\text{bf}} \) can be exploited for early anticipation of \( \langle k_{\text{final}} \rangle \) and therefore of the ratio of better-fit/less-fit sites in the final clone.

Connection between the Genetic Distance and the Variance of the Fitness Distribution

The total congruence \( C_{\text{tot}} \) correlates closely with the variance of the distribution \( f(k, t) \). Figure 2b shows that a proportionality of the form

\[
\text{Var}[k] = \frac{L^{\frac{1}{2}}}{2}(1 - C_{\text{int}})
\]

holds with excellent accuracy over the entire range of population sizes, with \( \alpha \) a constant coefficient close to unit, \( \alpha = 0.85 - 0.97 \). We note that \( L(1 - C_{\text{tot}}) \) represents the average Hamming distance within the population. Relation (11) implies, then, that the rate of evolution, equation (3), is proportional to the average Hamming distance, as

\[
\frac{d\langle k \rangle}{dt} = -\alpha \frac{L^{\frac{1}{2}}}{2}(1 - C_{\text{int}}).
\]

The origin of equation (11) for \( \alpha = 1 \) can be understood under the assumption that the distribution of less-fit alleles on polymorphic sites within any pair of genomes is random, as it is the case, for example, when nonfixed alleles are randomly distributed in the population. To this end consider the following expression for the variance \( \text{Var}[k] \):

\[
\text{Var}[k] = \langle (k - \langle k \rangle)^2 \rangle = \frac{1}{2N^2} \sum_{i,j=1}^{N} (k_i - k_j)^2,
\]

where \( i \) and \( j \) label individual genomes. For each pair \((i, j)\), let \( n_{\text{poly}} \) denote the number of polymorphic sites between the respective genomes and let \( n_{\text{f}} \) be the corresponding number of nonmonomorphic mutant sites, so that \( k_i + k_j = n_{\text{poly}} + 2n_{\text{f}} \). Then each term \( (k_i - k_j)^2 \) can be rewritten as \( 4(\Delta k_i - n_{\text{poly}}/2)^2 \), where \( \Delta k_i = k_i - n_{\text{f}} \) denotes the number of polymorphic mutant sites in genome \( i \). But because mutations in pairs of polymorphic sites are distributed with equal chance (1/2) between the 2 sites, \( n_{\text{poly}}/2 \) is just the average number of polymorphic mutations that can be contributed by genome \( i \) for fixed \( n_{\text{poly}} \) and one has \( (k_i - k_j)^2 = 4(\Delta k_i - \Delta k_j)^2 \). Further, if \( \Delta k_i \) does not correlate between genomes, for example, less-fit alleles for a fixed number of polymorphic sites are distributed randomly
between the 2 genomes, from the binomial distribution we have, on average, \[ \frac{\Delta k_i}{C_{tot}} \sim 2 = n_{poly}/4. \] By definition \( n_{poly} = L(1 - C_{tot}(i,j)) \), and from equation (13), it follows that \( \text{Var}[k] = \frac{L}{2}(1 - C_{tot}) \). Correlations between the numbers \( \Delta k_i \) of polymorphic mutations, associated with correlations between fitness values of different genomes [Rouzine and Coffin 2005], lead to \( (\Delta k_i - \Delta k_j)^2 < \Delta k_i/2 \), which accounts for \( \alpha < 1 \) in equation (11). We note that in the limit of large \( N \), when \( \frac{L}{2}(1 - C_{tot}) \sim (k)(1 - \frac{(k)}{T}) \), coefficient \( \alpha \) approaches the ratio \( p = s_{eff}/s \), as defined in Rouzine and Coffin (2005), \( \alpha \approx p \), whereas at lower \( N \), equation (11) still holds, but \( \alpha \neq p \) (see fig. 2b).

**Linkage Disequilibrium**

The initial population has a random allele makeup, with null pairwise linkage disequilibrium as defined by equation (10). Representative tests of pairwise linkage disequilibrium at later times are shown in figure 3 for 2 typical
sets of model parameters and various linkage distances between sites of a pair, \( d \). The quantity \( D'(d) \), calculated according to equation (10), exhibits only small, statistically insignificant fluctuations from a null value, irrespective of the distance \( d \). In particular, this result shows that the positions of alleles of a specific type (beneficial or deleterious) do not correlate along genomes, although alleles (either deleterious or beneficial) do correlate between homologous sites on different genomes (across the population), as described in the previous paragraph. We note in this regard that the experimental observation of linkage disequilibrium can be attributed to random mutation events that generate new beneficial alleles (Rouzine and Coffin 1999), as well as to epistasis. In the present model, epistasis is absent, and all the alleles are assumed to preexist.

Dependence of the Evolution Rate and the Sequence Congruence on \( N \) and Other Parameters

As documented in figure 2b, the analytic approach based on the effective selection coefficient \( s_{\text{eff}} \), equation (6), provides a good fit to the simulated results for the kinetics of \( \text{Var}[k] \) and \( \langle k \rangle \). As can be seen in figure 4, for a fixed characteristic population size \( N_0 \) and raw selection coefficient \( s \), the average over all runs of the ratio \( s_{\text{eff}}/s \) shows a logarithmic increase with the population size \( N \) and reaches an upper plateau at the deterministic value of 1 for large enough \( N \). In this limit, recombination is so effective that sites that are statistically independent at the onset evolve independently henceforth. At lower population sizes, \( s_{\text{eff}}/s \) also shows a tendency toward a plateau. Closer examination reveals that the onset of this lower plateau correlates with the gradual disappearance of simulation runs that end in the population with maximum fitness, \( \langle k \rangle = 0 \) (i.e., for a small enough population size, the final clones start to include with certainty some less-fit alleles). To emphasize this feature, we calculated, in addition, a fractional ratio \( s'_{\text{eff}}/s \), defined as the contribution to the average \( s_{\text{eff}}/s \) arising exclusively from the fraction of runs arriving at the best-fit population (fig. 4). Whereas at sufficiently large \( N \) the fractional ratio \( s'_{\text{eff}}/s \) is identical to \( s_{\text{eff}}/s \), at the lower population sizes, where runs that end in the best-fit population gradually disappear, the fraction displays an abrupt drop and reaches a null value at some critical population size \( N_c \). This behavior may be regarded as a phase transition from all-best-fit final clones to partially better-fit final clones.

We observe also that the increase of \( s_{\text{eff}}/s \) with population size \( N \) is closely matched by a similar decrease in sequence congruence, as given by \( s_{\text{eff}}/s \sim 1 - C_{\text{min}}^{\text{tot}} \) (fig. 4). Furthermore, the ratio \( s_{\text{eff}}/2 \) of the evolution rate to the average Hamming distance \( L(1 - C_{\text{tot}}) \) in equation (12) depends only weakly on the population size (result not shown).

Finally, we note that our results do not greatly depend on the number of crossovers, \( M \), per genome per recombination event. Although the majority of our simulations used a value \( M = 10 \), we verified that the dependence of \( s_{\text{eff}} \) on \( N \)
does not change much in the limit $M \to \infty$. For instance, for a rather high recombination rate, $N_0 = 10^3$ and long genomes, $L = 1000$, an increase in the number of crossovers produced a shift in the critical population size from $N_c \approx 300$, for $M = 10$, to $N_c \approx 200$, for $M \to \infty$.

**Discussion**

**Evolution Rate versus Accumulation of Sequence Correlation**

A previous approach (Rouzine and Coffin 2005) to the present selection/recombination model considered only correlations between the number $k$ of less-fit alleles in different genomes but neglected allele correlations at homologous sites. Under this approximation, it was found that the rate of evolution was described by the effective selection coefficient approach, equations (3 and 5a), with coefficient $s_{\text{eff}}$ given by equation (5b). The ratio $s_{\text{eff}}/s$ was found to have a threshold in the population size at $N_c \approx N_0^{1/2}$, equation (5b), below which evolution stopped shortly after its onset $(s_{\text{eff}}/s=0)$.

The present Monte Carlo results confirm the existence of a threshold in $N$ for $s_{\text{eff}}/s$ (fig. 4). Also, for fixed $N_0$ and $s$, the fractional ratio $s_{\text{eff}}/s$ shows an increase with population size. There are, however, a number of differences between the previous approximation and the present simulations. First, the transition between $s_{\text{eff}}=0$ and $s_{\text{eff}}=s$ occurs in a more narrow interval of $\ln N$. Second, the value of the critical population size $N_c$ is higher than $\approx N_0^{1/2}$. Third, the current simulations indicate that the accumulation of pairwise congruence, in excess of what one would expect from a uniformly random distribution of alleles, plays an essential role in the evolution of a viral population (fig. 2a). At the same time, pairwise linkage disequilibrium between sites within the same genome does not emerge (fig. 3).

A qualitative understanding of correlation buildup can be gained on noticing that 2 genomes must correlate if they share segments originating in an identical ancestral clone. To quantify this effect, let us introduce the “identity by descent” $C$ of a pair of sequences, as the fraction of sites that have a common ancestor, that is, are identical by descent. Two homologous sites in a pair of sequences descend from either 2 different genomes in the initial population or the same genome. In the second case, a site belongs to one of either 2 different genomes in the initial population or the same genome. In the second case, a site belongs to one of 2) was tailored for a quantitative understanding of long-term HIV

and

$$1 - C_{\text{tot}} = (1 - C)(1 - f^2 - (1 - f)^2). \quad (14c)$$

In our model, we expect $C$ to increase monotonously in time, from a vanishing value, $C(t=0)=0$, for the uncorrelated initial population at $f=1$, to some finite value $C(t=\infty) \leq 1$, for the final clone at $f = 0$. With this conjecture, it can be shown from equation (12) that the sequence congruences will have most of the qualitative features (monotony, minima, asymptotic trend at $f = 0$ and $f = 1$) displayed in figure 2a.

One can use equation (14) to compare values of the identity by descent $C$ obtained in 3 independent ways. In figure 2a, we show the 3 values of $C$ calculated in this way and numerical values for $C_{\text{eff}}, C_{\text{bf}}$, and $C_{\text{tot}}$ for a parameter value set near the transition point in $N$, where $s_{\text{eff}}/s \approx 0.5$. We observe that 1) the 3 curves for $C$ are close, which confirms the assumption that sites not identical by descent do not correlate and 2) the average value of $C$ in the interval of $f$ is not small, demonstrating the connection between sequence correlation and the kinetic properties of population. We note that, at small $f$, $C_{\text{bf}}$ can be used as a crude approximation for $C$, equation (14a).

The effective selection coefficient can be expressed in terms of the identity by descent $C$. It is sufficient to note that sites identical by descent (and therefore having identical alleles) do not contribute, on average, to the variance $\text{Var}[k]$ of the number of deleterious mutations because $\text{Var}[k]=1/2N^2 \sum_{i,j}(k_i - k_j)^2$. Hence, for a population size $N$ sufficiently large to avoid any loss of better-fit alleles and neglecting correlation in $k$ between genomes (Rouzine and Coffin 2005), one can write

$$\text{Var}[k] \approx (1 - C)(k) \left( 1 - \frac{k}{L} \right). \quad (15)$$

Note that equations (15 and 14c) produce equation (12) for $\alpha = 1$. This interpretation is essentially based on the assumption that partial correlation between 2 genomes at individual sites does not occur: sites either have a common ancestor or do not correlate at all. In this regime, equation (15) and equation (5a) allow one to identify

$$\frac{s_{\text{eff}}}{s} \approx 1 - C. \quad (14)$$

Furthermore, as the population size increases, so does the number of clones represented in the population, whereas the frequency of genomes belonging to any given clone decreases. This relationship means that the accumulation of identity by descent decreases, and therefore, $C$ decreases monotonously with $N$. Thus, equation (14) implies, in turn, that $s_{\text{eff}}$ increases monotonously with $N$, in agreement with our numerical observations.

**Relevance for HIV Populations**

Our choice of numerical values for the model parameters ($L = 10–100, N < 10^3, N_0 = 10^3$ to $10^5, s = 10^{-2}$ to $10^{-1}$, initial frequency of better-fit alleles $\approx 10^{-2}$) was tailored for a quantitative understanding of long-term HIV
evolution, particularly for the reversion of deleterious alleles in untreated low-viremia patients. Our results are also relevant for improvement of the existing antiviral therapy targeted at delay of emergence of drug-resistant variants. In this case, the initial state corresponds to the small virus population surviving drug depletion and harboring a very low number of better-fit, drug-resistant alleles (frequency $\sim 10^{-2}$). The subsequent evolution at constant population size models the initial period of accumulation of drug-resistant alleles at a low viral load, before the onset of uncontrolled rebound.

Current drug cocktails that achieve $>99.99\%$ virus depletion target a small number of sites in HIV genome. The level of resistance (fitness increase) sufficient for virus replication under drug is not evident from the level of depletion and has to be estimated from a population dynamics model. A recent work of one of us (Rouzine et al. 2006) estimates that the decrease of the wild-type virus fitness due to drug by a factor of 10–20 may be sufficient to achieve such a deep level of depletion. Because $L \sim 2–3$ mutations enable virus to replicate effectively under drug, the selection coefficient under therapy is large for these critical mutations, $s = \ln(10–20)/L = 0.7–1.5$. Thus, the present simulations, using large $L$ and small $s$, do not apply to evolution of the critical resistance mutations under the current therapy regimens. As shown in Rouzine and Coffin (2005) using a separate argument, in this case, resistant mutations become fixed in the surviving population within a few generations, unless the effective population size is extremely low, $N_c < 30–45$.

On the other hand, both the analytic results (Rouzine and Coffin 2005) and the present simulations suggest that such drastic depletion levels are not necessary for successful control of resistance development. According to our model, if the number of drug target sites is increased to, for example, $L = 10$ (which somewhat exceeds the number of currently known drug targets), we predict that the rebound of resistant strains will require much larger population sizes. The selection coefficient per critical resistant site, in this case, is smaller, $s = \ln(10–20)/L = 0.2–0.3$, and the present theory applies approximately. Our previous estimate of the critical population size is $N_c \approx N_0^{1/2}$ (Rouzine and Coffin 2005). The present numerical results, figure 4, show that the actual critical population size, below which completely resistant strains do not materialize with certainty, is considerably larger, as given by the empirical relationship $N_c \sim (15–30) N_0^{1/2}$ for $N_0 = 10^3$ to $10^5$, $L = 10–100$, and $s = 0.1$. Using the only available estimate of recombination rate $r = N/N_0 \sim 1$ (Jung et al. 2002) and the estimate of population size in an average patient $N \sim 10^6$ (Rouzine and Coffin 1999), we obtain $N_0 \sim 10^3$ and $N_c \sim 10^4$. The value of $r$ may be overestimated because most coinfecting virions may originate from nearby cells and be genetically uniform. The estimate of $N$ also has to be updated to account for a higher number of crossovers than thought previously and for the multisite nature of evolution (the quoted estimate was based on a 2-site model). However, even assuming that the value of $N_0$ is 2 orders of magnitude smaller, that is, $N_0 = 10^2$, we still obtain $N_c \sim 1000$ (fig. 4), much larger than the critical depletion level for currently used drug cocktails, $N_c \sim 30$.

In summary, our current results strongly reinforce the conclusion set forth in the previous study (Rouzine and Coffin 2005) that successful long-term antiretroviral therapy could be achieved with significantly reduced dosage, provided the number of target sites is sufficiently large.

Acknowledgments

This work was supported by grants K25 AI 01811 to I.R. and R01 CA 089441 to J.M.C. S.G. was supported by training grant T32 AI007389. J.M.C. was an American Cancer Society Research Professor with support from the George Kirby Foundation.

Literature Cited


Pekka Pamilo, Associate Editor

Accepted November 21, 2006