Recombination increases human immunodeficiency virus fitness, but not necessarily diversity

N. N. V. Vijay,1 Vasantika,1 Rahul Ajmani,2 Alan S. Perelson2 and Narendra M. Dixit1

1Department of Chemical Engineering, Indian Institute of Science, Bangalore 560012, India
2Theoretical Biology and Biophysics, Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Recombination can facilitate the accumulation of mutations and accelerate the emergence of resistance to current antiretroviral therapies for human immunodeficiency virus (HIV) infection. Yet, since recombination can also dissociate favourable combinations of mutations, the benefit of recombination to HIV remains in question. The confounding effects of mutation, multiple infections of cells, random genetic drift and fitness selection that underlie HIV evolution render the influence of recombination difficult to unravel. We developed computer simulations that mimic the genomic diversification of HIV within an infected individual and elucidate the influence of recombination. We find, interestingly, that when the effective population size of HIV is small, recombination increases both the diversity and the mean fitness of the viral population. When the effective population size is large, recombination increases viral fitness but decreases diversity. In effect, recombination enhances (lowers) the likelihood of the existence of multi-drug resistant strains of HIV in infected individuals prior to the onset of therapy when the effective population size is small (large). Our simulations are consistent with several recent experimental observations, including the evolution of HIV diversity and divergence in vivo. The intriguing dependencies on the effective population size appear due to the subtle interplay of drift, selection and epistasis, which we discuss in the light of modern population genetics theories. Current estimates of the effective population size of HIV have large discrepancies. Our simulations present an avenue for accurate determination of the effective population size of HIV in vivo and facilitate establishment of the benefit of recombination to HIV.

INTRODUCTION

The identification of improved treatment protocols and the development of new, more potent drugs and vaccines for human immunodeficiency virus (HIV) infection hinge on our understanding of the mechanism(s) that underlie the ability of HIV to diversify its genomic content and develop multi-drug resistance during current antiretroviral therapies and/or escape vaccine-induced cell-mediated immune responses (Barouch et al., 2002; Clavel & Hance, 2004; Simon et al., 2006; Thomson et al., 2002). Several competing factors influence HIV diversification in infected individuals: the high mutation rate of HIV, approximately \(3 \times 10^{-5}\) mutations per nucleotide per replication (Mansky & Temin, 1995), and the enormous production rate, approximately \(10^{10}\) virions per day in a chronically infected individual (Perelson et al., 1996), enable rapid sampling of broad spectra of genomic variations by HIV. Consequently, viral genomes with drug resistance mutations may exist in individuals prior to the onset of therapy (Lech et al., 1996; Ribeiro & Bonhoeffer, 2000). On the other hand, random genetic drift, significant in the suggested small effective cell populations in vivo (Shriner et al., 2004b), and fitness selection (Kils-Hutten et al., 2001) may impede HIV diversification. Experiments reveal complex patterns of the genomic evolution of HIV in patients and suggest that the patterns are correlated with disease progression (Herbeck et al., 2006; Markham et al., 1998; Shankarappa et al., 1999). Several recent studies present insights into the complex interplay of mutation, drift and fitness selection that underlies HIV diversification in vivo (Frost et al., 2000, 2001; Rouzine et al., 2001; Williamson et al., 2005; Yuste et al., 2002). The role of recombination, however, remains poorly understood (Bocharov et al., 2005; Bretschler et al., 2004; Fraser, 2005; Otto & Lenormand, 2002).

HIV has a high recombination rate, several times larger than its point mutation rate (Jetzt et al., 2000; Levy et al., 2004; Shriner et al., 2004a; Suryavanshi & Dixit, 2007). Further, recent studies suggest that infected cells harbour multiple HIV proviruses (Chen et al., 2005; Dang et al., 2004; Jung et al., 2002; Levy et al., 2004), which enables the
formation of heterozygous virions and presents the necessary substrate for recombination to induce genomic diversification (Rhodes et al., 2003). Recombination may thus significantly accelerate viral diversification (Bocharov et al., 2005; Charpentier et al., 2006) and facilitate the emergence of multi-drug resistance in infected individuals (Althaus & Bonhoeffer, 2005; Blackard et al., 2002; Christiansen et al., 1998; Kellam & Larder, 1995; Moutouh et al., 1996).

Conversely, just as recombination may induce the accumulation of favourable mutations, it may also drive beneficial combinations of mutations apart and therefore not necessarily benefit HIV (Bonhoeffer et al., 2004; Fraser, 2005). Mathematical models (Bretcher et al., 2004; Fraser, 2005) suggest that the effect of recombination depends sensitively on epistasis ($E$), which is a measure of the influence of interactions between mutations on viral fitness (in a two-locus-two-allele model, where haplotype $ab$ has fitness $f_{ab} = f_{AB} - f_{Ab} + f_{aB} - f_{ab}$). In particular, when drug-sensitive loci interact antagonistically in reducing viral fitness, i.e. when $E > 0$, recombination may decelerate the emergence of multi-drug resistance in a large population of cells (Bretcher et al., 2004). Recent experiments suggest that the HIV-1 fitness landscape is characterized by mean $E > 0$, raising doubts on the benefits of recombination to HIV (Bonhoeffer et al., 2004). With moderate values of the effective population size, $N_e$, recombination may accelerate the emergence of drug resistance significantly even with $E > 0$ (Althaus & Bonhoeffer, 2005). When $N_e$ is smaller than a critical value, however, the viral population in an individual may converge to a clone in the absence of mutation, leaving little opportunity for recombination to induce diversification (Rouzine & Coffin, 2005). Current estimates of $N_e$ in vivo have large discrepancies (Kouyos et al., 2006a) and leave unclear the benefit of recombination to HIV.

Our aim is to develop a fundamental understanding of the effect of recombination on HIV diversification in an infected individual. Mathematical modelling of HIV diversification requires integration of the effects of mutation, multiple infections of cells, random genetic drift, fitness selection, and recombination, which remains difficult especially when fitness interactions between multiple loci are important. Conversely, experimental approaches suffer from the difficulty involved in the deconvolution of the effects of recombination from those of mutation, drift and selection. Here, we develop computer simulations that elucidate the influence of recombination on the genomic diversification of HIV.

**METHODS**

We performed bit-string simulations of the within-host genomic evolution of HIV (Fig. 1). We commenced simulations with a pool of identical homozygous virions that synchronously infect a pool of uninfected cells. Following infection, viral RNAs in cells are reverse-transcribed into proviral DNA. During reverse transcription, mutation and recombination introduce genomic variations. The proviral DNAs are transcribed into viral RNAs, which are assorted randomly into pairs, copackaged and released as new virions. The progeny virions form a new viral pool from which virions are chosen according to their fitness to infect a new pool of uninfected cells, and the replication cycle is repeated. In each generation, we determine the average diversity and fitness of the proviral DNA and their average divergence from the proviral DNAs at the start of infection. Below, we present details of the simulation procedure.

**Fig. 1.** A schematic representation of the simulation procedure.
Initialization of the viral pool. We consider a pool of $N_v$ virions. Each virion is a string of $2L$ nucleotides (A, G, C and U); the first $L$ nucleotides represent one RNA strand and the next $L$ nucleotides the other RNA strand of the virion. At the start of infection, a sequence, $F$, of length $L$ is generated by randomly selecting a nucleotide (from A, G, C and U) at each position. The sequence is assumed to be the fittest and is assigned the relative fitness 1. We mutate $F$ with probability $\xi$ per nucleotide (see below) and let the resulting sequence be the starting or founder sequence. The founder sequence is copied to the two substrings of length $L$ in each of the $N_v$ virions to form the initial viral pool.

Infection of cells. We let $M$ virions infect each cell in a pool of $N_v$ cells. Virions are chosen for infection from the viral pool according to their relative fitness (see below) and their genomes are transferred to cells.

Reverse transcription. Following infection, reverse transcription converts the RNA strands (length $2L$) to proviral DNA (length $L$). Reverse transcription involves recombination, which we simulate first, and mutation (Bocharov et al., 2005).

Recombination. The average recombination rate, $\rho$, is approximately $8 \times 10^{-4}$ per site per reverse transcription (Levy et al., 2004; Suryavanshi & Dixit, 2007) so that for the genome lengths we consider ($L=40$–$100$), more than one crossover per reverse transcription is unlikely. Thus, during each reverse transcription, we allow a single crossover with probability $pL$. The crossover occurs at position $l_i$, which we choose randomly from a uniform distribution on 1 to $L$. We begin reverse transcription from the first nucleotide of one of the two substrings of the viral RNA chosen randomly with probability 0.5. We copy the nucleotide sequence of the starting substring from the first to the $l_i$th position to the first $l_i$ positions of the resulting proviral DNA. The template is then switched and the sequence from the $(l_i+1)$th position to the last position of the other substring is copied to the corresponding positions in the proviral DNA. This process is repeated for all infecting virions.

Mutation. Mutations in HIV are dominated by substitutions of which approximately 90% are transitions (A→G and C→T) (Mansky & Temin, 1995). For simplicity, we therefore consider transitions alone. Thus, starting from the first position of a proviral DNA, ‘A’ is substituted with ‘G’ and ‘C’ with ‘T’, and vice versa, at all positions, each substitution occurring with probability equal to the mutation rate, $\mu$.

Calculation of diversity, divergence and fitness. The normalized Hamming distance between two proviral DNAs, $i$ and $j$, is

$$d_{ij} = \frac{1}{L} \sum_{k=1}^{L} \left[ 1 - \delta(x_i^k - x_j^k) \right]$$

where $x_i^k$ is the nucleotide in the $k$th position of provirus $i$, and $\delta(x_i^k = x_j^k) = 1$ if $x_i^k = x_j^k$ and $\delta(x_i^k \neq x_j^k) = 0$ if $x_i^k \neq x_j^k$ (2). $d_{ij}$ is thus the number of differences per position between sequences $i$ and $j$; $d_{ij} = 0$ if the two strings are identical, whereas $d_{ij} = 1$ if the two strings are different at every position.

We determine the average diversity of the $Q$ (=MN_v) proviruses in any generation as

$$d_G = \frac{1}{Q(Q-1)} \sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} d_{ij}$$

where the Hamming distance between every possible pair of proviruses is averaged. The denominator, $(1/2)Q(Q-1)$, is the number of possible pairs among $Q$ proviruses. Similarly, the average divergence of the proviruses in any generation from those at the onset of infection, is

$$d_5 = \frac{1}{Q} \sum_{i=1}^{Q} d_0$$

where $d_0$ is the Hamming distance of provirus $i$ in a given generation from the founder sequence (subscript 0).

The fitness landscape for the protease and reverse transcriptase regions of HIV-1 has recently been determined (Bonhoeffer et al., 2004). On this landscape, the relative fitness, $f_i$ of genome $i$ depends on the Hamming distance, $d_{ij}$, between the amino acid sequence of genome $i$ and that of the fittest genome, $F$. As an approximation, we assume that unit Hamming distance between two nucleotide sequences is equal to unit Hamming distance between the corresponding amino acid sequences. We find then that the experimental fitness landscape is captured by (Supplementary Fig. S1, available with the online version of this paper),

$$f_i = 1 - (1 - f_{\text{min}})^{(d_{0i})^n}$$

where $f_{\text{min}}=0.269$ is the minimum fitness of sequences attained at arbitrarily large absolute Hamming distances from $F$; $d_{0i}L=20.4$ is that Hamming distance from $F$ at which $f_i=0.5(1+f_{\text{min}})$, i.e. the average of the fittest and the least fit sequences; and $n=3$ is analogous to the Hill coefficient. The average fitness of proviruses in any generation is

$$f_{\text{avg}} = \frac{1}{Q} \sum_{i=1}^{Q} f_i$$

Translation, assortment and viral production. Following reverse transcription, each infected cell produces $P$ virions. For each progeny virion arising from a cell, two of the $M$ proviral DNAs from the cell are chosen at random and their sequences are copied to the two $L$-bit substrings to form the RNA genome of the virion. When $M=1$, the same proviral sequence is copied to all the new virions arising from that cell. The new virions thus formed constitute the viral pool for infecting the next generation of target cells.

Fitness selection. The relative fitness of a virion is determined by equation 5 with $d_{ij}$ the average Hamming distance of its two RNA strands from the fittest sequence. The virion is chosen for infection randomly with a probability equal to its relative fitness. The process is repeated for other virions in the pool until every cell is infected with $M$ virions.

The above infection and replication process is repeated and the evolution of viral diversity, divergence and fitness with the number of replication cycles determined. Several realizations are averaged to determine the expected evolution. The simulations are implemented using a computer program written in C++.

RESULTS

We begin simulations with the minimal processes essential for viral diversification and gradually build complexity until the influence of recombination is elucidated. We then perform simulations with parameter values representative of conditions in vivo and draw links with patient data.

Mutation

We examine first the effect of mutation. We consider a viral pool of $N_v=40$ virions, each virion with a genome of
L = 40 nucleotides. We assume individual cells to be infected by single virions (M = 1) and let each infected cell produce a single progeny virion (P = 1). The cell pool contains Nc = 40 cells. We ignore recombination (r = 0) and assume a flat fitness landscape (fi = 1).

In Fig. 2, we present the evolution of diversity, dG, and divergence, dS (Methods), with the number of replication cycles or generations, η, for different values of the mutation rate, μ. As expected, dG = dS = 0 when η = 0, as all the virions in the initial viral pool are identical. As η increases, both dG and dS increase due to mutation. The rate of increase is higher for larger values of μ. For large η, both dG and dS approach equilibrium values, dG' and dS', respectively, in an exponential manner; we find that dG' = dS' = 0.5 independent of μ.

We capture these effects of mutation in a mathematical model (Supplementary Information, available with the online version of this paper). Model predictions are in excellent agreement with our simulations (Fig. 2) and provide a quantitative understanding of the influence of mutation on HIV diversification as well as a successful test of our simulation procedure.

Recombination (r > 0) does not alter the genomic evolution in Fig. 2 because, in the absence of multiple infections, all progeny virions are homozygous and no substrate exists for recombination to induce diversification (Supplementary Fig. S2). We therefore introduce multiple infections of cells next.

Multiple infections of cells

With Nc = 40 cells, we expand the viral pool to Nv = NcM so that every cell is infected by M virions, and let P = M, so that all progeny virions infect cells. We choose μ = 0.001 substitutions per nucleotide per replication, r = 0 and fi = 1.

In Fig. 3(a), we present the evolution of dG for different values of M = P. For M = 1, the evolution is identical to that in Fig. 2. For M = 2, surprisingly, dG decreases significantly from that for M = 1. The decrease in dG despite multiple infections is attributed to enhanced viral production from cells. When P = 2, each cell produces two progeny virions, which can be identical. The production of identical virions lowers the diversity of the proviruses in the following generation, and, in consequence, dG. This effect is

---

**Fig. 2.** The evolution of HIV diversity (a) and divergence (b) at different mutation rates, viz., 0.05 (+), 0.01 (○), 0.005 (○), 0.001 (△), 0.0005 (□), and 0.0001 (○) substitutions per nucleotide per replication. The grey bands represent deviations to the mean (± SD) due to stochastic variations in approximately 1000 realizations. Solid lines are model predictions (Supplementary Information).

**Fig. 3.** The evolution of HIV diversity at different frequencies of multiple infections of cells (M) and numbers of progeny virions per cell (P). (a) M = P = 1 (△), 2 (○), 3 (□), 4 (○); (b) M = 1; P = 2 (○), 3 (□), 4 (○); (c) P = 3; M = 1 (○), 2 (□), 3 (○).
emphasized in Fig. 3(b), where we present \( d_G \) for \( M=1 \) and \( P=2, 3 \) and 4. As \( P \) increases, the probability that identical virions infect target cells increases and causes a decrease in \( d_G \). We note, however, that when \( P>M \), the decrease in \( d_G \) is the combined effect of the production of identical virions and random genetic drift, as only a fraction of the progeny virions infects cells. The effect of drift is minimized when \( P=M \).

As \( M(=P) \) increases above 2, \( d_G \) increases (Fig. 3a). Enhanced multiple infections increase the likelihood of diverse proviruses forming progeny virions and cause the observed increase in \( d_G \). This phenomenon is illustrated further in Fig. 3(c), where we present the evolution of \( d_G \) for \( M=1, 2 \) and 3, but for fixed \( P=3 \). As \( M \) increases, \( d_G \) increases.

Multiple infections (\( M>1 \)) and drift (\( P>M \)) do not influence HIV divergence; in the absence of fitness selection and recombination, the evolution of individual sequences is governed by mutation. \( d_k \) therefore remains unaltered by variations in \( M \) and \( P \) (Supplementary Fig. S3).

Even with multiple infections, recombination (\( \rho>0 \)) does not alter HIV diversification (Supplementary Fig. S4). Whereas recombination can bring mutations together, it can also drive mutations apart. Thus, when no fitness benefit exists for accumulating (or separating) mutations, i.e. when no non-random association of mutations is favoured, HIV diversification remains independent of recombination. We examine next the role of fitness selection.

**Fitness selection**

To delineate the effect of fitness selection, we ignore multiple infections (\( M=1 \)) and select virions for infection according to equation 5 from a pool of size \( N_v=N_fP \) with \( P>1 \). Without loss of generality, we let the founder sequence be the fittest (\( \zeta=0 \)) (Methods). The fitness of individual genomes is thus a function of their divergence.

In Fig. 4, we present the evolution of \( d_G, d_0 \) and the average fitness, \( f_{\text{avg}} \) for \( P=3 \) and different values of \( \mu \) (when \( \rho=0 \)). We find that fitness selection influences both \( d_G \) and \( d_0 \), \( d_G^* \) and \( d_0^* \) decrease compared with their values with a flat fitness landscape (Figs 2, 3 and 4). The relative fitness of genomes decreases as their divergence increases (equation 5). Thus, fitness selection lowers \( d_G^* \) (Fig. 4b). The fewer variations allowed due to fitness penalties also lower \( d_G^* \) correspondingly (Fig. 4a). A higher \( \mu \), however, forces the sampling of larger genomic variations and causes \( d_G^* \) and \( d_0^* \) to increase (Fig. 4). Correspondingly, the equilibrium fitness, \( f_{\text{avg}}^* \) decreases as \( \mu \) increases (Fig. 4c), in accordance with the classical mutation–selection balance (Hartl & Clark, 2007).

The evolution of \( d_G \) is more complex (Fig. 4a). For small and large values of \( \mu, d_G \) increases monotonically to \( d_G^* \). For intermediate values of \( \mu \), the evolution is non-monotonic: \( d_G^* \) first rises and then declines to \( d_G^{**} \) (Fig. 4a inset). To understand this non-monotonic evolution, we recognize that, as \( d_G \) increases with \( \eta \) (Fig. 4b), the average number of mutations per genome increases. Initially (\( \eta \sim 0 \)), when all genomes are nearly identical, mutations tend to occur at distinct positions on different strands, causing \( d_G \) to rise. As the number of mutations accrued increases, the likelihood that new mutations occur at novel positions decreases. Eventually, new mutations occur increasingly at positions where other strands carry
mutations, causing $d_G$ to decline until the mutation–selection balance is attained. When $\mu$ is small, the maximum number of mutations accrued is so small that no reduction in $d_G$ occurs, whereas when $\mu$ is large, forceful diversification outweighs fitness selection and results in the monotonic evolution of $d_G$ similar to that in the absence of fitness selection (Fig. 2).

Recombination does not alter the HIV diversification observed in Fig. 4 as expected from the lack of multiple infections of cells and the consequent absence of heterozygous virions (Supplementary Fig. S5).

**Recombination**

With multiple infections of cells and fitness selection, recombination influences HIV diversification. In Fig. 5(a–c), we present the evolution of $d_G$, $d_S$, and $f_{avg}$ with $M=2$ and fitness selection according to equation 5 for different recombination rates, $\rho$. Remarkably, we find that recombination does not alter the initial rate of evolution of $d_G$, $d_S$ or $f_{avg}$, but alters $d_G^\infty$, $d_S^\infty$ and $f_{avg}^\infty$. As $\rho$ increases, $d_G^\infty$ and $f_{avg}^\infty$ increase and $d_S^\infty$ decreases (Fig. 5a–c). These qualitative effects of recombination are robust to variations in the mutation rate and quantitative variations in the fitness landscape (i.e. changes in $d_S$ or $f_{min}$ in equation 5) (not shown), but are sensitive to the effective population size.

**Effective population size**

Upon increasing the cell population, $N_c$, at any value of $\rho$, $d_G^\infty$ increases because a larger number of genomic variants is likely to exist in larger populations. At the same time, $f_{avg}^\infty$ increases, and hence $d_S^\infty$ decreases, because fitness selection operates more effectively in larger populations (Fig. 5). Remarkably, however, as $N_c$ increases, $d_G^\infty$ first increases upon increasing $\rho$ (Fig. 5a and d) and then decreases (Fig. 5g). The difference in $d_G^\infty$ which we denote $\Delta d_G^\infty$ upon increasing $\rho$ from 0 to 0.02 crossovers per nucleotide per replication cycle, decreases from approximately 0.015 at $N_c=40$ to approximately $-0.028$ at $N_c=5000$ (Fig. 6a). Thus, recombination increases HIV diversity for small $N_c$ but decreases diversity for large $N_c$. The corresponding difference in the fitness (divergence), $\Delta f_{avg}^\infty$, is positive (negative) for all values of $N_c$ considered (Fig. 6a), indicating that recombination consistently increases (lowers) the mean fitness (divergence) of the viral population. The extent of the increase in fitness, however, decreases upon increasing $N_c$. We discuss these intriguing results in the light of modern population genetics theories below.

**In vivo parameter estimates**

To determine whether the above effects of recombination may hold in vivo, we perform simulations with parameter values that mimic conditions in vivo. We let $\mu=3 \times 10^{-5}$.

---

**Fig. 5.** The evolution of HIV diversity (top), divergence (middle), and average fitness (bottom panels) at different recombination rates, viz., 0 (red), 0.0002 (orange), 0.002 (blue) and 0.02 (green) crossovers per nucleotide per replication, and at different sizes of the cell pool, viz., 40 (a–c), 200 (d–f) and 1000 (g–i) cells. Mutations occur at 0.001 substitutions per nucleotide per replication and fitness selection follows equation 5. Each cell is infected by $M=2$ virions and produces $P=3$ progeny virions.
substitutions per position per replication (Mansky & Temin, 1995) and \( M = 3 \) (Jung et al., 2002). The viral burst size \textit{in vivo} is estimated to be in the range \( 10^5 - 10^6 \) (Haase et al., 1996; Hockett et al., 1999). Most virions produced, however, may be non-infectious (Dimitrov & Martin, 1995; Dimitrov et al., 1993; Piatak et al., 1993) and individual cells may yield as few as 2-3 infectious virions (Dimitrov & Martin, 1995). Here, we therefore choose \( P = 5 \). We assume fitness selection to follow equation 5. We vary \( \rho \) from 0 to \( 9 \times 10^{-3} \) crossovers per position per replication, approximately tenfold higher than the experimental recombination rate (Levy et al., 2004), to emphasize the influence of recombination. The best current estimates of the effective population size are of the order of \( 10^3 \) (Shriner et al., 2004b). We therefore vary \( N_e \) from 10 to \( 10^4 \). We also consider a larger genome length, \( L = 100 \), corresponding to the variable portion of the V1V2 region of \textit{env} (Bocharov et al., 2005). With \( L = 100 \), we span the experimental fitness landscape, which extends to a Hamming distance of approximately 60 (Supplementary Fig. S1).

In Fig. 6(b), we present the resulting effect of recombination and \( N_e \) on \( d_c^s, d_S^c \) and \( f_{avg} \). In agreement with our observations above (Fig. 6a), we find that, for small values of \( N_e \), recombination increases both the mean fitness and the diversity of the viral genomes, whereas, for large values of \( N_e \), recombination increases the mean fitness but lowers viral diversity (Fig. 6b).

**Links with patient data**

We demonstrate finally that the predicted evolution of diversity and divergence are consistent with corresponding observations in HIV patients. In Fig. 7, we compare our simulations with the evolution following seroconversion of the diversity and the divergence of the C2–V5 region of the HIV-1 \textit{env} gene in one (Patient 2) of nine patients observed experimentally (Shankarappa et al., 1999). The patient data are reported as the mean pairwise distance between viral DNA sequences determined using either a two-parameter Kimura model or a general time-reversible model with site-to-site variation in substitution rates, both methods yielding similar results (Shankarappa et al., 1999). When the distances are small, as is the case in the experimental data, the Kimura two-parameter distance reduces to the Hamming distance (Kimura, 1980), allowing us to compare our simulation results directly with the data.

We choose the above \textit{in vivo} parameter estimates, fix \( \rho = 9 \times 10^{-4} \) crossovers per position per replication (Levy et al., 2004), and let the viral generation time be 1 day (Dixit et al., 2004; Markowitz et al., 2003; Perelson et al., 1996; Rodrigo et al., 1999). \( \xi \), which determines the fitness of the founder sequence, and \( N_e \) for the patient remain unknown. We vary these parameters and find that good comparisons between our simulations and the patient data are obtained for \( N_e = 1500 \), \( \xi = 0.05 - 0.08 \) and \( N_e = 5000 \), \( \xi = 0.05 - 0.06 \) (Fig. 7). In the patient, viral diversity rises following seroconversion, appears to attain a maximum and then declines slightly to an equilibrium value. With \( \xi = 0.05 \) and \( N_e = 1500 \) or 5000, our simulations capture accurately the initial rise and the equilibrium diversity (Fig. 7a and b). The evolution of diversity, however, is monotonic. Given the large uncertainties associated with the experimental data, whether the maximum observed is significant remains to be ascertained.

Our simulations with the same parameter values also capture the evolution of viral divergence in the patient (Fig. 7c and d). We find here that the experimental data are encompassed by our predictions with \( \xi = 0.05 \) and 0.08 for \( N_e = 1500 \) and \( \xi = 0.05 \) and 0.06 for \( N_e = 5000 \). We assume a unique founder sequence in our simulations. In the patient, however, even if infected with a unique founder, a mixture of genomes, perhaps with fitness values in the range determined by the above ranges of \( \xi \), may exist at seroconversion. Note that variation of \( \xi \) has little impact on the evolution of diversity (Fig. 7a and b).

The reasonable agreement between the data and our simulations (Fig. 7) suggests that our simulations are
representative of the scenario in vivo. The agreement also suggests that $N_e$ in the patient considered is in the approximate range of 1500–5000 cells. [Simulations with $N_e=500$ underpredict and $N_e=10000$ overpredict the patient data (Fig. 7). Note that $\xi$ does not influence the equilibrium diversity or fitness but alters equilibrium divergence (Supplementary Information). Also note that the error bars in Fig. 7 are extreme values and not standard errors of the mean.] With $N_e \sim 1500–5000$, our simulations predict that recombination enhances mean viral fitness but lowers viral diversity in the patient (Fig. 6b). Whether recombination exerts a similar influence in other patients remains to be ascertained.

DISCUSSION

Our simulations are in agreement with several recent studies of the effects of recombination on HIV evolution. Resistance mutations in HIV-1 appear to exhibit mean positive epistasis ($E>0$) (Bonhoeffer et al., 2004). With $E>0$, the reduction due to recombination of the mean time for the establishment of double mutants is predicted to increase, reach a maximum, and then decrease as the effective population size, $N_e$, increases (Althaus & Bonhoeffer, 2005). In our simulations, the equilibrium diversity, $d^*_G$, provides a qualitative measure of the waiting time for the establishment of drug resistance. The larger the value of $d^*_G$, the greater is the likelihood of the existence of diverse genomes in the viral pool, and hence the shorter is the waiting time for the establishment of desired mutants.

Our simulations employ the fitness landscape identified experimentally (Bonhoeffer et al., 2004). Indeed, we find that recombination increases $d^*_G$ when $N_e$ is small and decreases $d^*_G$ when $N_e$ is large (Fig. 6).

Recent computer simulations suggest that recombination and multiple infections act synergistically in the evolution of HIV diversity (Bocharov et al., 2005). With multiple infections, the mean diversity and fitness of HIV increase in the presence of recombination. Our simulations are in agreement with these predictions and provide further insights. We note first that recombination enhances both viral fitness and diversity when $N_e$ is small. Second, an increase in the frequency of multiple infections ($M$) increases diversity when viral production per cell ($P$) is relatively unaffected. If $P$ increases with $M$, which may happen when viral production is limited by a viral rather than a host-cellular factor, then HIV diversity may first decrease and then increase as the frequency of multiple infections increases (Fig. 3).

In vivo, recombination allows the conservation of genomic regions affected by evolutionary bottlenecks and increases diversity in other regions (Charpentier et al., 2006). Thus, where selection dominates, as with large $N_e$, recombination facilitates selection of fitter genomes, and where selection is weak, recombination increases diversity, as observed in our simulations (Figs 5 and 6).

The patterns of change of viral diversity, $d_G$ and divergence, $d_S$, predicted by our simulations have been observed in experiments (Herbeck et al., 2006; Markham et
The observed patterns may also arise due to HIV-mediated collapse of the immune system, i.e. immune relaxation (Williamson et al., 2005), which we ignore. We have applied our simulations to describe the time-evolution of $d_C$ and $d_G$ in one HIV patient for a period of approximately 10 years following seroconversion. Our simulations capture the data well (Fig. 7), suggesting that our simulations are representative of the scenario in vivo.

Current population genetics theories provide insights into the influence of recombination elucidated by our simulations (Ewens, 2004; Hartl & Clark, 2007; Kouyos et al., 2007; Otto & Lenormand, 2002; Rice, 2002). According to these theories, recombination acts to reduce the magnitude of linkage disequilibrium (LD). In a two-locus-two-allele model, where $g_{ab}$ is the frequency of haplotype $ab$, $\text{LD}=g_{ab}g_{AB}−g_{AB}g_{ab}$ measures the extent to which the frequency of double-mutants is different from that expected from the frequencies of single-mutants. In the absence of selection and with large $N_e$ LD vanishes. Recombination then will not influence viral diversification (Supplementary Figs S3–S5).

Random genetic drift generates negative LD – which implies an overrepresentation of single-mutants – according to the well-known Hill–Robertson effect (Hill & Robertson, 1966). Epistasis also introduces LD; then LD has the same sign as $E$ (Eshel & Feldman, 1970). When $N_e$ is small, drift may dominate epistasis and create net negative LD. When recombination lowers negative LD, the frequency of double-mutants and hence viral diversity increases (Hartl & Clark, 2007; Kouyos et al., 2007; Otto & Lenormand, 2002). When $N_e$ is large, drift is reduced. LD is then determined by epistasis. In our simulations, we have assumed a fitness landscape that extends over multiple loci and has mean $E>0$ (Bonhoeffer et al., 2004). Although LD depends in a complicated manner on the distribution of epistasis and not the mean epistasis alone (Kouyos et al., 2006b), the assumed landscape may be expected to generate positive LD. When recombination lowers positive LD, viral diversity is expected to decrease (Hartl & Clark, 2007; Kouyos et al., 2007; Otto & Lenormand, 2002). Thus, a competition between negative LD introduced by drift and positive LD due to epistasis appears to underlie the effect of recombination and $N_e$ on viral diversity observed in our simulations.

Our simulations also suggest that for small $N_e$ ($<10^5$), the enhancement in viral diversity due to recombination increases with increasing $N_e$ (Fig. 6b). For very small $N_e$, the Hill–Robertson effect may not apply (Otto & Barton, 2001). Thus, the gradual emergence of the Hill–Robertson effect with increasing, yet small, $N_e$ may underlie the observed enhanced effect of recombination. We observe finally that in all our simulations, recombination increases mean viral fitness (Fig. 6) and hence decreases the mutational load, an effect suggested as one of the plausible causes of the evolutionary origins of recombination (Hartl & Clark, 2007; Kouyos et al., 2007; Otto & Lenormand, 2002).

Prediction of the influence of recombination on viral diversification and consequently the emergence of drug resistance has been precluded by the lack of robust estimates of $N_e$ in vivo (Kouyos et al., 2006a). Models based on neutral evolution estimate $N_e$ to be approximately $10^5$ (Achaz et al., 2004; Brown, 1997; Nijhuis et al., 1998; Rodrigo et al., 1999; Seo et al., 2002; Shriner et al., 2004b). HIV evolution in vivo, however, is expected to be driven by selection. A model that assumes viral evolution with selection estimates $N_e$ to be $>10^5$–$10^6$ (Rouzine & Coffin, 1999). The model, however, is restricted to fitness interactions between pairs of loci and ignores recombination, which may overestimate $N_e$ (Shriner et al., 2004b). In our simulations, we consider fitness interactions between multiple loci and predict the evolution of viral diversity and divergence. The evolution of viral diversity and divergence are sensitive to $N_e$. Comparison of our predictions with data of the evolution of viral diversity and divergence from patients therefore presents an avenue for obtaining more accurate estimates of $N_e$ in vivo. We demonstrate the applicability of this methodology by analysing data from one patient and estimate $N_e$ to be approximately in the range 1500–5000. Analysis of data from a larger set of patients – a promising potential application of our simulations – would provide robust estimates of $N_e$ in vivo and establish the benefit of recombination to HIV.

**ACKNOWLEDGEMENTS**

Portions of this work were done under the auspices of the US Department of Energy under contract DE-AC52-06NA25396. This work was supported by the National Institutes of Health grants AI065334 (NMD), AI28433 and RR06555 (A. S. P.), and the NIH-funded Center for HIV/AIDS Vaccine Immunology.

**REFERENCES**


[http://vir.sgmjournals.org](http://vir.sgmjournals.org)
state of knowledge and potential clinical consequences. Clin Infect Dis 34, 1108–1114.


