Quantifying Selection against Synonymous Mutations in HIV-1 *env* Evolution

Fabio Zanini, Richard A. Neher

Evolutionary Dynamics and Biophysics Group and Max Planck Institute for Developmental Biology, Tübingen, Germany

Intrapatient evolution of human immunodeficiency virus type 1 (HIV-1) is driven by the adaptive immune system resulting in rapid change of HIV-1 proteins. When cytotoxic CD8$^+$ T cells or neutralizing antibodies target a new epitope, the virus often escapes via nonsynonymous mutations that impair recognition. Synonymous mutations do not affect this interplay and are often assumed to be neutral. We test this assumption by tracking synonymous mutations in longitudinal intrapatient data from the C2-V5 part of the *env* gene. We find that most synonymous variants are lost even though they often reach high frequencies in the viral population, suggesting a cost to the virus. Using published data from SHAPE (selective 2′-hydroxyl acylation analyzed by primer extension) assays, we find that synonymous mutations that disrupt base pairs in RNA stems flanking the variable loops of gp120 are more likely to be lost than other synonymous changes: these RNA hairpins might be important for HIV-1. Computational modeling indicates that, to be consistent with the data, a large fraction of synonymous mutations in this genomic region need to be deleterious with a cost on the order of 0.002 per day. This weak selection against synonymous substitutions does not result in a strong pattern of conservation in cross-sectional data but slows down the rate of evolution considerably. Our findings are consistent with the notion that large-scale patterns of RNA structure are functionally relevant, whereas the precise base pairing pattern is not.

Human immunodeficiency virus type 1 (HIV-1) evolves rapidly within a single host during the course of the infection. This evolution is driven by strong selection imposed by the host immune system via cytotoxic CD8$^+$ T cells (CTLs) and neutralizing antibodies (nAbs) [1] and is facilitated by HIV-1's high mutation rate [2, 3]. Escape mutations in epitopes targeted by CTLs are typically observed during early infection and spread rapidly through the population [4]. During chronic infection, the most rapidly evolving parts of the HIV-1 genome are the variable loops (V1 to V5) in the envelope protein gp120 (V loops), which change to avoid recognition by nAbs. Escape mutations in *env*, the gene encoding gp120, spread through the viral population within a few months. Consistent with this time scale, it is found that serum from a particular time typically neutralizes autologous virus extracted more than 3 to 6 months earlier but not contemporary virus [5].

Escape mutations are selected because they change the amino acid sequences of viral proteins in a way that reduces antibody binding or epitope presentation. Conversely, synonymous mutations do not modify the viral proteins and are commonly used as approximately neutral markers in studies of viral evolution, i.e., as a negative control for detecting selected sites [6–8]. In addition to maintaining protein function and avoiding the adaptive immune recognition, however, the HIV-1 genome has to ensure efficient processing and translation, nuclear export, and packaging into the viral capsid: all these processes operate at the RNA level and are sensitive to synonymous changes. Several important RNA secondary structures have been characterized in detail, including the HIV-1 Rev response element (RRE) in *env* which enhances nuclear export of full-length or partially spliced viral transcripts via a complex hairpin RNA structure [9]. In fact, the HIV-1 genome is full of RNA structures [10] with no or unknown function. However, large-scale modification of secondary structures can result in substantial reduction of the replication capacity [11], and the propensity of forming RNA stems anticorrelate with the rate of evo-

lution [12, 13]. These poorly characterized RNA structures are conserved to different degrees in HIV-1 and simian immunodeficiency virus (SIV): corresponding regions tend to be part of similar structural elements, but individual base pairings are very rarely conserved [14].

In this paper, we characterize the dynamics of synonymous mutations in *env* and show that, in the region of the V loops, a large fraction of these mutations are deleterious. Despite their fitness cost, deleterious synonymous variants rise in frequency in the viral population via genetic hitchhiking due to limited recombination in HIV-1 populations [15, 16]. We show a strong correlation between the fate of a synonymous variant and the surrounding RNA structure. We then compare our observations to computational models and obtain estimates for the effect of synonymous mutations on viral fitness.

MATERIALS AND METHODS

Sequence data collection. Longitudinal intrapatient viral RNA sequences were collected from published studies [17–19] and downloaded from the Los Alamos National Laboratory (LANL) HIV sequence database [20]. The viral RNA sequences from some patients show substantial population structure and were excluded (see Fig. S1 in the supplemental material); a total of 11 patients with 4 to 23 time points each and approximately 10 sequences per time point were analyzed. The time intervals between two

consecutive sequences ranged from 1 to 34 months, with most of them between 6 and 10 months.

Sequence analysis. The sequences were translated, and the resulting amino acid sequences were aligned to each other and the NL4-3 reference sequence separately for each patient, using MUSCLE (21). For the sequences from each patient, the consensus nucleotide sequence at the first time point was used to classify alleles as "ancestral" or "derived" at all sites. Sites with high frequencies of gaps were excluded from the analysis to avoid artifactual substitutions due to alignment errors. Allele frequencies at different time points were extracted from the multiple-sequence alignment.

A mutation was considered synonymous if it did not change the amino acid corresponding to the codon and if the rest of the codon was in the ancestral state. Codons with more than one mutation were discarded. Slightly different criteria for synonymous/nonsynonymous discrimination yielded similar results.

Fixation probability and secondary structure. For the estimates of time to fixation/extinction, single nucleotide variants (SNVs) were binned by frequency, and the time to first fixation or extinction was stored. The fixation probability was determined as the long-time limit of the resulting curves. Mutations that reached high frequency but neither fixed nor were lost were classified as "floating" unless they first reached high frequencies within 3 years of the last time point. In that case, it was assumed they had not had sufficient time to fix and were discarded.

The SHAPE (selective 2′-hydroxyl acylation analyzed by primer extension) scores quantifying the degree of base pairing of individual sites in the HIV-1 genome were downloaded from the journal website (10). Wherever possible, SHAPE reactivities were assigned to sites in the multiple-sequence alignments for each patient through the alignment to the sequence of the NL4-3 virus used in reference 10. Problematic assignments in indel-rich regions were excluded from the analysis. The variable loops and flanking regions were identified manually starting from the annotated reference HXB2 sequence from the LANL HIV database (20).

Computer simulations. Computer simulations were performed using FFPopSim (22). Briefly, FFPopSim enables individual-based simulations where each site in the genome is represented by one bit that can be in one of two states. Outcrossing rates, crossover rates, mutation rates, and arbitrary fitness functions can be specified. While the best estimate for the HIV generation time is roughly 2 days (23, 24), the generation time has very little influence on the results and basically sets the unit of time by which other parameters are measured. For simplicity, we used one generation per day and have checked that the results obtained are indistinguishable from simulations with a generation time of 2 days.

Parameters. In order to simulate HIV evolution we have to specify several parameters that are known to various degrees. The mutation rate has been measured using exogenous LacZ constructs, and the average nucleotide substitution rate is estimated to be $\mu \approx 2 \times 10^{-5}$ per generation (2, 3). In our simulations, we vary the mutation rate μ from 10^{-5} to 4×10^{-5} per nucleotide per day. Different virions within an infected person can recombine their genome by coinfection of the same cell followed by template switching during reverse transcription. Template switching happens around 10 times per reverse transcription (25). The combined recombination rate of template switching and coinfection has been estimated based on modeling to be around $\rho \approx 10^{-5}$ per nucleotide per day (15, 16), implying a coinfection rate of a few percent, as has recently been confirmed experimentally (26). We assume a template switching rate of 10^{-3} per nucleotide and vary the recombination rate between 5×10^{-6} to 5×10^{-5} per nucleotide per day by adjusting the coinfection rate. The population size relevant to evolution in chronic infection has been hotly debated. The number of cells infected by HIV-1 per day is on the order of 10^7 (23), but the number of viruses contributing to the next generation might be considerably smaller due to the burstiness of replication. Recent evidence points to a relevant population size in excess of the inverse mutation rate (27). In fact, the evolutionary dynamics depends only weakly on the actual value of the population size N once

the product $N\mu$ is of order one or larger (28). We simulate population sizes between 10^4 and 5×10^4 virions. Larger populations become prohibitively costly to simulate.

Another crucial ingredient for our simulations is the fitness landscape, i.e., the effects of mutations on fitness and possible interactions between mutations. For simplicity, the third positions of every codon were deemed synonymous and assigned either a selection coefficient of 0 with probability α or, with probability $1 - \alpha$, a deleterious effect of magnitude s_d. In our simulations, α varies from 0 to 0.75 and s_d varies from 2×10^{-4} to 2×10^{-2} per day. Mutations at the first and second positions were assigned deleterious fitness effects of -0.02 (simulations with a larger cost of -0.2 were also performed and produced similar results). With a probability of k_A per generation, a random locus in the genome is designated an epitope that can escape by one or several mutations with exponentially distributed escape rates. We denote the mean escape rate by ε. The rate of escape in chronic infection is reported to be on the order of a few percent (29, 30), consistent with the finding that virus is neutralized by serum from 3 to 6 months later (5). We simulated escape rates between $2 \times 10^{-2.5}$ and $2 \times 10^{-1.5}$ per day. The rate k_A at which new antibody challenges arise is more difficult to quantify, but a substantial fraction of nonsynonymous substitutions in gp120 are probably driven by escape. We simulate a range from $10^{-3.8}$ to $10^{-1.5}$ per day.

For the models with competition within epitopes, a complex epistatic fitness landscape was designed such that each single mutant is sufficient for full escape. Specifically, each mutation has an additive effect equal to the escape rate but interacts with all other escape mutations in the same epitope with a negative effect of the same magnitude. Higher-order terms were included to make sure that not only double mutants but all multiple mutants had the same fitness (see supplemental material). To model recognition of escape variants by the evolving immune system, the beneficial effect of an escape mutation was set to its previous cost of -0.02 with a probability per generation proportional to the frequency of the escape variant.

Sampling and analysis. To obtain reasonable sampling of a particular parameter combination, we ran simulation for 6,000 days, and we repeated each run 100 times with different seeds for the random number generator. Both full-length HIV-1 genomes and *env*-only simulations were performed and yielded comparable results. Populations were initialized with a homogeneous founder population. After 30 generations of burn-in to create genetic diversity, new epitopes were introduced at random with rate k_A. The simulations were repeated 3,000 times with a seven-dimensional Latin hypercube sampling scheme (31) bounded by the ranges given for each of these parameters above. For all parameters except the fraction α of neutral third-position sites, parameters were sampled uniformly in log space.

The areas below or above the neutral fixation probability (diagonal line) were estimated from the binned fixation probabilities using linear interpolation between the bin centers. The bins used were [0.05, 0.15], [0.15, 0.25], [0.25, 0.4], and [0.4, 0.6], which is sufficiently precise for our purposes. Note that we did not consider the frequency interval between 0.6 and 1, because very few variants are observed in this window and its inclusion generates more noise than signal.

Methods availability. All analysis and computer simulation scripts, as well as the sequence alignments used, are available for download at http://git.tuebingen.mpg.de/synmut.

RESULTS

Due to the large population size and the high mutation rate, every possible single nucleotide variant (SNV) is produced multiple times per day (32). Some of these variants rise to high enough frequency that they are observed in a sample of sequences. SNVs rise or fall in frequency because of three reasons: (i) their own effect on fitness and escape, (ii) their association to genetic backgrounds, and (iii) stochastic fluctuations (genetic drift). We study the dynamics of sample frequencies, ν, of SNVs, i.e., the fraction ν
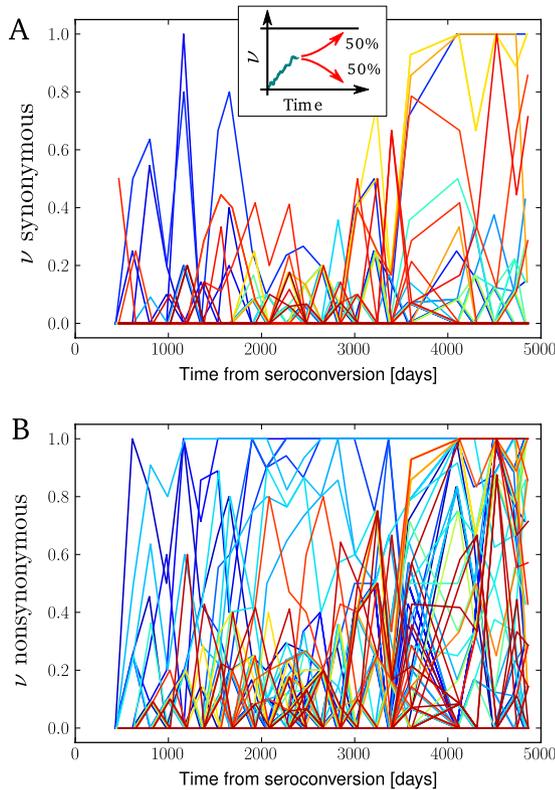
FIG 1 Time series of frequencies of synonymous (A) and nonsynonymous (B) single nucleotide variants (SNVs) in *env*, C2-V5, from patient p10 (17). While many nonsynonymous SNVs fix, few synonymous SNVs do so even though they are frequently observed at high frequencies. Colors indicate the position of the site along the C2-V5 region (blue to red). (Inset) The fixation probability P_{fix} of a neutral SNV that reached 50% frequency is one half.
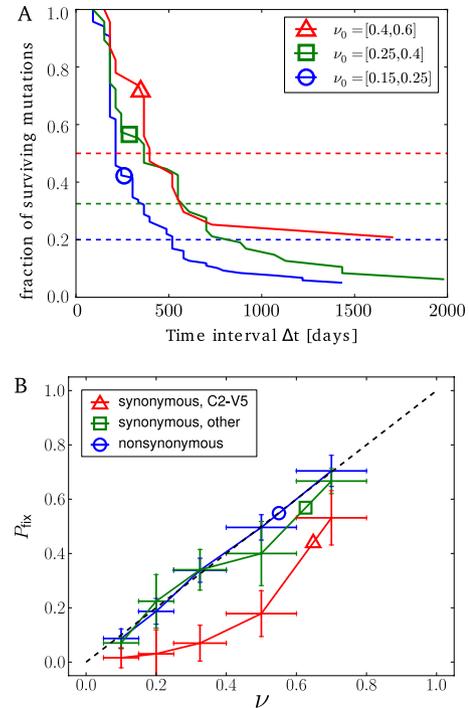

FIG 2 Fixation and loss of SNVs. Panel A shows how quickly synonymous SNVs are purged from the populations. Specifically, the figure shows the fraction of SNVs that are still observed after Δt days, conditional on being observed in one of the three frequency intervals (which are shown in different colors). In each frequency interval, the fraction of synonymous SNVs that ultimately survive is the fixation probability P_{fix} conditional on the initial frequency. The neutral expectation for $P_{fix} = \nu_0$ is indicated by dashed horizontal lines. Panel B shows the fixation probability of synonymous SNVs as a function of ν_0. Polymorphisms within C2-V5 fix less often than expected for neutral SNVs (indicated by the black dashed diagonal line). This suppression is not observed in other parts of *env* or for nonsynonymous SNVs. The horizontal error bars indicate the bin sizes, and the vertical ones indicate the standard deviation after 100 patient bootstraps of the data. Data in this figure are taken from references 17, 18, and 19.

of the sequences in a sample carrying the variant. When an SNV is present in all sequences at a certain time point, we say it has "fixed"; when it is completely absent, we say it was "lost" or is "extinct."

Most positions are only transiently variable, and variants will either fix or will be lost—at least in small samples. Given that an SNV is at a certain frequency ν, the probability of fixation is higher for beneficial SNVs than for neutral ones; in turn, neutral variants fix more frequently than deleterious ones. The fixation probability of a neutral SNV at frequency ν is the frequency itself, i.e., $P_{fix}(\nu) = \nu$, while it goes extinct with probability $1 - \nu$. For instance, if a neutral SNV is observed in half of the sequences, it will fix with a probability of 50% (see inset in Fig. 1A). The fixation probability of neutral SNVs is independent of most model assumptions and is affected only if neutral SNVs are associated preferentially with viruses with either high or low fitness.

Figure 1 shows the time course of the frequencies of all synonymous SNVs (Fig. 1A) and nonsynonymous SNVs (Fig. 1B) observed in the C2-V5 region of *env* in a chronically HIV-1-infected patient (p10 from Shankarappa et al. [17]). Despite many synonymous SNVs reaching high frequency, very few fix (Fig. 1A); in contrast, many nonsynonymous mutations fix (Fig. 1B). This observation seems at odds with the assumption of neutrality.

Many synonymous SNVs in C2-V5 are deleterious. We studied the dynamics and fate of synonymous variants more quantitatively by analyzing data from seven patients from Shankarappa et

al. (17) and Liu et al. (18) as well as three patients from Bunnik et al. (19) (patients with viruses with strong viral population structure were not considered; see Materials and Methods and Fig. S1 in the supplemental material). The former data set is restricted to the C2-V5 region of *env*, while the data from Bunnik et al. (19) cover most of *env*. We considered all SNVs in a frequency interval $[\nu_0 - \delta\nu, \nu_0 + \delta\nu]$ at some time t and calculated the fraction that is still observed at later times $t + \Delta t$. Plotting this fraction against the time interval Δt, we see that most synonymous SNVs segregate for roughly 1 year and are lost much more frequently than expected under neutrality (Fig. 2A). The long-time probability of fixation, P_{fix}, is shown as a function of the initial frequency ν_0 in Fig. 2B. The neutral expectation is shown as a black dashed line. We found that P_{fix} of synonymous variants is far below the neutral expectation in C2-V5 (red line). Outside C2-V5, using data from Bunnik et al. (19) only, we found no such reduction in P_{fix} (green line). Restricted to the C2-V5 region, the sequence samples from Bunnik et al. (19) are fully compatible with data from Shankarappa et al. (17) and show that synonymous mutations fix less often than expected under neutrality. The nonsynonymous SNVs seem to follow more or less the neutral expectation (blue line)—a point to which we return below.
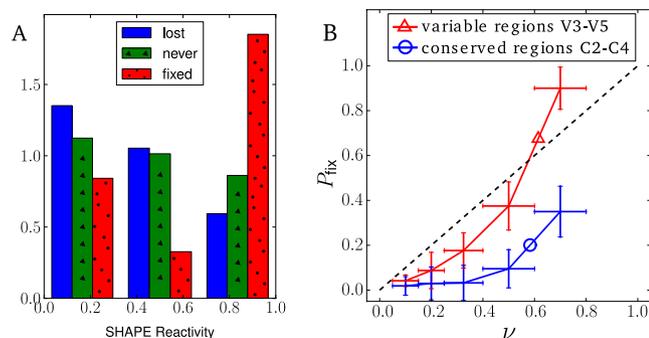
FIG 3 Permissible synonymous mutations tend to be unpaired. (A) Distribution of SHAPE reactivities among sites at which synonymous SNVs fixed, sites at which SNVs reached frequencies above 15% but were subsequently lost, and sites at which high-frequency SNVs were never observed (all categories are restricted to the regions V1-V5 [within 100 bp of V1-V5]). Sites at which SNVs fixed tend to have higher SHAPE reactivities, corresponding to less base pairing, than those at which SNVs are lost. Sites at which no SNVs were observed show an intermediate distribution of SHAPE values. (B) Fixation probability of synonymous SNVs in C2-V5 separately for variable regions V3-V5 and the connecting conserved regions C2-C4 that harbor RNA stems. As expected, the fixation probability is lower inside the conserved regions. Data in this figure are taken from references 10, 17, 18, and 19.

When interpreting these results for the fixation probabilities, it is important to note that we focused on SNVs that have already reached high frequencies. In HIV-1 infection, most SNVs remain very rare throughout: they are not considered here. Synonymous SNVs can reach high frequencies either through genetic drift or genetic hitchhiking on escape variants (see below); very deleterious variants will never reach high frequencies in the first place. Hence, our analysis indicates that, among all synonymous SNVs that somehow reach high frequencies, most of those in C2-V5 are deleterious, while those in the rest of *env* tend to be neutral.

Synonymous mutations in C2-V5 tend to disrupt RNA stems. One possible explanation for a reduced fixation of synonymous variants in C2-V5 is secondary structure in the viral RNA, the disruption of which is deleterious to the virus (12, 13, 33).

The propensity of nucleotides in the HIV-1 genome to form base pairs has been measured using the SHAPE assay, a biochemical reaction preferentially altering unpaired bases (10). The SHAPE assay has shown that the V1 to V5 variable regions tend to be unpaired, while the conserved regions between these stretches form stems. We aligned the sequences from each patient to the reference NL4-3 strain used by Watts et al. (10) and assigned SHAPE reactivities to most positions in the alignment. We then calculated the distributions of SHAPE reactivities for synonymous SNVs that fixed or were lost (only variants reaching frequencies above 15%). As shown in Fig. 3A, the reactivities of fixed SNVs (red histogram) are systematically larger than those of lost SNVs (blue) (Kolmogorov-Smirnov test on the cumulative distribution, $P \approx 0.002$). In other words, SNVs that are likely to break RNA helices are also more likely to revert and finally be lost from the population, restoring the helix. Note that this analysis will be sensitive only at positions where the base pairing pattern of NL4-3 agrees with that of each patient's initial consensus sequence—it is thus statistically conservative. For a control, we also calculate the distribution of SHAPE reactivities for SNVs that never reach high frequencies (green). This set is a mixture of neutral and deleterious SNVs, and as expected, its distribution lies between those of fixed and lost high-frequency variants.

To test the hypothesis that synonymous variants in C2-V5 are lost because they break stems in the conserved stretches between the V loops, we considered SNVs in V loops and their flanks separately. The greatest depression in fixation probability is observed in the conserved stems, while the V loops show little deviation from the neutral signature (Fig. 3B). This is suggestive of important RNA helices in conserved regions between the V loops.

In addition to RNA secondary structure, we have considered other possible explanations for a fitness cost of some synonymous mutations, in particular codon usage bias (CUB). HIV-1 is known to prefer A-rich codons over highly expressed human codons (34, 35). We did not find any evidence for a contribution of average CUB to the ultimate fate of synonymous SNVs; this is consistent with the observation that HIV-1 is not adapting its codon usage to its human host cells at the macroevolutionary level (35).

Deleterious SNVs can reach high frequency by hitchhiking. While the observation that some synonymous variants are deleterious is not unexpected, it seems odd that we observe them at high population frequency and that the fixation probability is reduced only in parts of the genome (in C2-V5 but not in the rest of *env*; compare the red and green lines in Fig. 2B). The C2-V5 region undergoes frequent adaptive changes to evade recognition by neutralizing antibodies (5, 36, 37). Due to the limited amount of recombination in HIV-1 (15, 16), deleterious variants that are linked to adaptive variants can reach high frequency. This process is known as hitchhiking (38) or genetic draft (28, 39). Hitchhiking is apparent in Fig. 1, which shows that many SNVs change rapidly in frequency as a cohort.

The approximate magnitude of the deleterious effects can be estimated from Fig. 2A, which shows the distribution of times after which synonymous SNVs at intermediate frequencies become fixed or lost. The typical time to loss is on the order of 500 days. If this loss is driven by the deleterious effect of the mutation, this corresponds to deleterious effects s_d of the order of 0.002 per day. (This is only an average estimate: each mutation is expected to have a slightly different fitness effect.)

To obtain a better idea of the range of parameters that are compatible with the observations and our interpretation, we performed computer simulations of a model of evolving viral populations assuming a mix of positive and purifying selection and rare recombination. For this purpose, we use the simulation package FFPopSim, which includes a module dedicated to intrapatient HIV evolution (22). For each simulation run, we specify the deleterious effect of synonymous mutations, the fraction of synonymous mutations that are neutral, the escape rate (selection coefficient) of adaptive nonsynonymous mutations, and the rate at which previously untargeted epitopes become targeted (the latter determines the number of sites available for escape). Note that the escape rate is the sum of two factors: the beneficial effect due to the ability to evade the immune system minus the fitness cost of the mutation in terms of its effects on structure, stability, etc. Net escape rates in chronic infections have been estimated to be on the order of $\varepsilon = 0.01$ per day (15, 29), consistent with a lag in neutralization of a few months (5).

Figure 4A shows simulation results for the fixation probability and the synonymous diversity for different deleterious effects or sizes of synonymous mutations. We quantify synonymous diversity via the fraction of sites with a synonymous SNV at a frequency $0.25 < \nu < 0.75$. We denote this fraction by P_{interm}. The synonymous diversity observed in patient data is indicated in the figure.
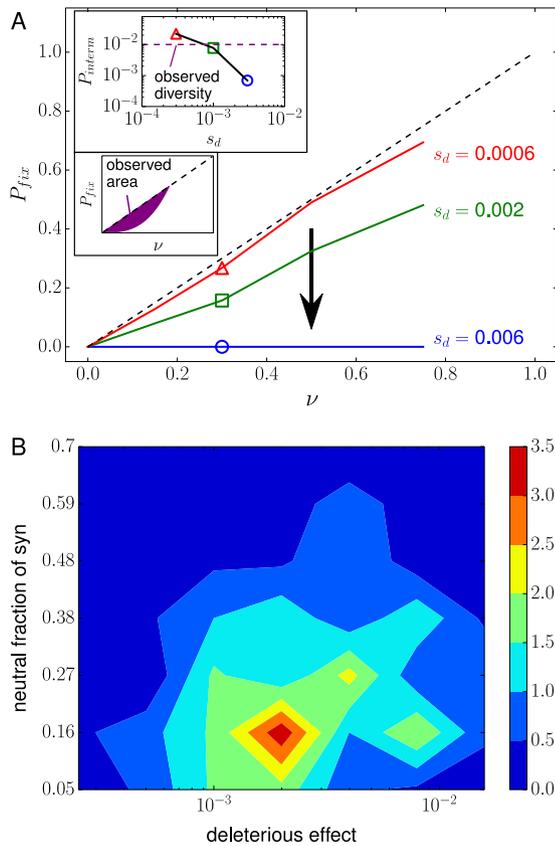
FIG 4 Distribution of selection coefficients on synonymous sites. (A) The depression in P_{fix} depends on the deleterious effect size of synonymous SNVs. This parameter also reduces synonymous diversity, measured by the probability of a SNV to be found at intermediate frequencies P_{interm} (top inset). The area observed in the data is also shown (bottom inset). (B) To assess the parameter space that affects synonymous fixation and diversity, we ran 3,000 simulations with random combinations of parameters for deleterious effect size, fraction of deleterious synonymous sites, average escape rate, rate of introduction of new epitopes, population size, mutation rate, and recombination rate (see Materials and Methods). The density of simulations that reproduce the synonymous diversity and fixation patterns observed in the data are shown. These simulations demonstrate that deleterious effects are around 0.002 and that a large fraction of the synonymous mutations need to be deleterious. The marginal distributions of each parameter are given in Fig. S5 in the supplemental material.

To quantify the depression of the fixation probability, we calculate the area between the measured fixation probability and the diagonal, which is the neutral expectation (Fig. 4A, bottom inset). We restrict this area up to frequency $\nu = 0.5$ because the high-frequency part is rather noisy. If no fixation happens, this restricted area will be -0.125; if every SNV fixes, the area will be 0.125. In data from HIV-1-infected patients, we find $P_{\text{interm}} \approx 0.01$, $A_{\text{syn}} \approx -0.09$ for synonymous changes and $A_{\text{nonsyn}} \approx 0$ for nonsynonymous changes. In the three simulations shown in Fig. 4A, the fixation probability of synonymous SNVs decreases from the neutral expectation ($A_{\text{syn}} = 0$) to zero ($A_{\text{syn}} = -0.125$) as the fitness cost of the SNVs increases; the synonymous diversity plummets as well, as deleterious SNVs are selected against.

To map the parameter range of the model that is compatible with the data, we repeatedly simulated the evolution for many different combinations of effects of synonymous mutations, s_d,

the fractions of synonymous sites that are neutral, α, escape rates and numbers of targeted epitopes. In addition, we varied the mutation rate, recombination rate, and population size (see Materials and Methods). Among all simulations, we selected the ones that show A_{syn} and P_{interm} as observed in the data, i.e., a large depression in fixation probability of synonymous SNVs but, simultaneously, a moderately high synonymous diversity. Figure 4B shows the distribution of deleterious effects s_d and neutral fraction α for which we found $-0.11 < A_{\text{syn}} < -0.06$ and $0.008 < P_{\text{interm}} < 0.015$. This subset of parameters indicates that in order to be consistent with the data, only a minority of synonymous sites ($\alpha \leq 0.2$) can be neutral and that the deleterious effects of the remainder are on the order of 0.002.

The observed distribution of α and s_d is plausible because of the following. (i) A substantial depression in P_{fix} requires pervasive deleterious SNVs, or the majority of SNVs reaching high frequency are neutral and no depression is observed. (ii) In order to hitchhike, the deleterious effect size has to be less than the escape rate. Otherwise, the double mutant (with both the escape mutation and the deleterious synonymous one) has little or no fitness advantage over the wild-type virus. (iii) The inverse of s_d is of the same order of magnitude as the fixation/extinction times estimated above (Fig. 2A). (iv) The cost s_d tends to be larger than 0.001, because mutations with smaller cost behave neutrally and often go to fixation as expected from the typical coalescent times observed for HIV-1.

The above simulations show that hitchhiking on favored nonsynonymous variants can reproduce the observed pattern of deleterious synonymous SNVs that rarely fix. While our model is already quite complicated with many poorly known parameters that make interpretation of simulation results challenging, we know that the actual evolutionary dynamics of virus within an infected person is much more complicated still. In our model, escape mutations are unconditionally beneficial and almost always fix once they reach high frequencies, i.e., A_{nonsyn} is well above zero. This is incompatible with the observed fixation probability of nonsynonymous SNVs that often disappear again (Fig. 2B). One possible reason for this discrepancy is the fact that the fitness landscape in our model is static: once a site is designated as targeted by the immune system, mutations at this site are beneficial forever. Upon inspection, the trajectories of nonsynonymous SNVs suggest the rapid rise and fall of many SNVs. Two possible mechanisms that could explain the transient rise of nonsynonymous SNVs are time-dependent selection and within-epitope competition.

If the immune system recognizes the escape mutant before its fixation, the mutant might cease to be beneficial and disappear soon, despite its quick initial rise in frequency. In support of this idea, Richman et al. (5), Wei et al. (37), and Bunnik et al. (19) report antibody responses against escape mutants. These responses are delayed by a few months, roughly matching the average time needed by an escape mutant to rise from low to high frequency. Furthermore, neutralization responses against early virus sometimes decline later in infection, which could lead to a reversion of some sites. To model this type of behavior, we assumed that antibody responses against escape SNVs arise at a rate proportional to the frequency of the escape variant and abolish the benefit of the escape mutations. As expected, this type of time-dependent selection retained the potential for hitchhiking but reduced fixation of nonsynonymous SNVs. Figure S3 in the supple-

mental material shows that P_{fix} of synonymous SNVs is not affected by this change, while P_{fix} of nonsynonymous SNVs approaches the diagonal as the rate of recognition of escape mutants is increased.

Furthermore, several different escape mutations might arise within the same epitope. Their benefits are not additive, because each mutation is essentially sufficient to escape and no additional benefit is gained from combining them. As a consequence, several escape SNVs can rise to high frequency rapidly; the one with the smallest cost in terms of replication, packaging, etc., is most likely to eventually fix, whereas all others are lost. The emergence of multiple competing escape variants within a single epitope has been shown for CTL and antibody escape (30, 37, 40, 41). This scenario has also been explicitly observed in the evolution of resistance to lamivudine (3TC), where the mutation M184V is often preceded by M184I (42). Consequently, there could be many nonsynonymous mutations that are beneficial only until the viral population has found a "better solution" and then subsequently disappear.

We implemented within-epitope competition in the model by allowing for multiple escape mutations per epitope that do not provide additional benefit to the virus when combined. Again, we found that the rapid rise of several nonsynonymous mutations to intermediate frequencies allows deleterious synonymous mutations to hitchhike, while the fixation probability of nonsynonymous mutations is reduced. With roughly six mutations per epitope, the simulation data are compatible with observations (see Fig. S4 in the supplemental material). The two scenarios, time-dependent selection and competition between equivalent escape pathways, are not exclusive, and possibly both are important in HIV-1 evolution. We note that our modeling merely suggests possible scenarios. More data are required to determine to what extent either of these mechanisms contributes to the observed pattern.

DISCUSSION

By analyzing the fate of single nucleotide variants (SNVs) in longitudinal data of HIV-1 *env* evolution, we demonstrated selection against synonymous substitutions in the relatively conserved regions C2-C4. We suggest through computational modeling that these SNVs have deleterious effects on the order of $s_d = 0.002$ per day and that the majority of all synonymous mutations in this region are deleterious. In the absence of hitchhiking in a large population, deleterious mutations with effect s_d should be at a frequency $\mu/s_d \approx 0.01$. The fact that we see many synonymous mutations at high frequency and that these SNVs nevertheless disappear over a time scale of a year suggests that they are brought to high frequency by hitchhiking on favorable genetic backgrounds. Although both the rise and fall of synonymous SNVs are subject to hitchhiking, the consistent loss of these variants conditional on being at high frequency implies selection against these SNVs.

Comparison with biochemical data (SHAPE) of base pairing propensity in the RNA genome of HIV-1 indicated that these mutations tend to disrupt RNA secondary structures (10). Computer models of RNA folding predict stable hairpins in these regions that have been suggested to be functional and termed "insulating stems" (10, 33). The weak selection against synonymous variants is compatible with the negative results of *in vitro* replication assays investigating the fitness effects of small RNA hairpins in HIV-1

(43): it would take hundreds of cell culture passages to detect fitness effects of the order of one per thousand. The longitudinal data, however, span many years, and our analysis is able to quantify the subtle fitness effect of RNA structure in intrapatient evolution. The fixation probabilities and the sojourn times of SNVs represent a rich and simple summary statistics of longitudinal sequence data. Most importantly, these statistics are informative even in the absence of a neutral control and are thus appropriate for analyzing properties of synonymous sites.

We find selection against synonymous substitutions despite the fact that the corresponding sites are not strongly conserved in cross-sectional data. This is consistent with a recent comparative analysis of SIV and HIV-1 RNA secondary structure using SHAPE assays and computational methods (14). While large-scale patterns of RNA structures tend to agree in both viruses, the individual base pairs forming the structures are almost always discordant. Even though the molecular architecture of these structures changes over time, selection seems to maintain them, which reduces the fixation probability and hence the rate of evolution at synonymous sites. As expected from this argument, the evolutionary rate at synonymous sites varies greatly along the HIV-1 genome (44) (see also Fig. S2 in the supplemental material). This variation can confound estimates of selection on proteins substantially (45). The dynamic nature of HIV secondary structure makes it difficult to find the compensatory mutations that would restore base pairing in the longitudinal data; in fact, the exact base pairing pattern is most likely different than in the reference sequence used for the SHAPE analysis. Nevertheless, the importance of secondary structure underscores the realization that viral fitness is a complicated function (fitness landscape) in the high-dimensional space of possible genotypes (46). We show that this landscape is shaped not only by interactions between amino acids but also by ubiquitous interactions between nucleotides in RNA structures.

Selection against the majority of synonymous substitutions is probably common across the genome, but we observe deleterious synonymous SNVs only at high frequency in C2-V5 of *env*, where they hitchhike to high frequencies on nAb escape mutations. Surprisingly, nonsynonymous mutations display a fixation pattern as if they were neutral (Fig. 2B). However, nonsynonymous diversity exceeds synonymous diversity despite the overall much greater constraints on the amino acid sequence, suggesting that the majority of high-frequency SNVs are escape mutations despite the fact that they are often lost again. We suggest that this paradoxical behavior could be due to (i) escape mutations that revert after they themselves are recognized by nAbs or (ii) the competition between different escape mutations within one epitope. Both mechanisms reduce the overall fixation probability and can give rise to the observed pattern of fixation in computer simulations. More data are required to investigate this behavior further.

The observed hitchhiking highlights the importance of linkage due to infrequent recombination for the evolution of HIV-1. The recombination rate has been estimated to be on the order of $\rho = 10^{-5}$ per base per day (15, 16, 47). It takes roughly $t = \varepsilon^{-1} |\log \nu_0|$ generations for an escape variant with escape rate ε to rise from an initially low frequency ν_0 to a frequency close to one. This implies that a region of length $l = (\rho t)^{-1} = \varepsilon/|\rho \log \nu_0|$ remains linked to the adaptive mutation. With $\varepsilon = 0.01$ and $\nu_0 \approx 10^{-3}$, we have $l \approx 100$ bases. Hence, we expect strong linkage between the variable loops and the flanking sequences but none far beyond the variable re-

gions, consistent with the lack of signal outside of C2-V5. In the case of much stronger selection—as is observed during early CTL escape or drug resistance evolution—the linked region is of course much larger (48).

While classical population genetics assumes that the dominant stochastic force is genetic drift, i.e., nonheritable fluctuations in offspring number, in large populations like HIV, stochasticity due to linked selection is much more important. Gillespie termed such fluctuations genetic draft (39), and their effect in facultatively sexual population such as HIV-1 has been characterized in reference 28. Importantly, large population sizes are compatible with low diversity and rapid coalescence when genetic draft dominates over genetic drift.

Our results emphasize the inadequacy of independent site models of HIV-1 evolution and the common assumption that selection is time independent or additive. If genetic variation is only transiently beneficial, existing estimates of the strength of selection (15, 16) could be substantial underestimates. Furthermore, weak conservation and time-dependent selection result in estimates of evolutionary rates that depend on the time interval of observation, with lower rates across larger intervals. This implies that deep nodes in phylogenies might be older than they appear.

ACKNOWLEDGMENTS

REFERENCES

1. Rambaut A, Posada D, Crandall KA, Holmes EC. 2004. The causes and consequences of HIV evolution. Nat. Rev. Genet. 5:52–61.
2. Mansky LM, Temin HM. 1995. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J. Virol. 69:5087–5094.
3. Abram ME, Ferris AL, Shao W, Alvord WG, Hughes SH. 2010. Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. J. Virol. 84:9864–9878.
4. McMichael AJ, Borrow P, Tomaras GD, Goonetilleke N, Haynes BF. 2010. The immune response during acute HIV-1 infection: clues for vaccine development. Nat. Rev. Immunol. 10:11–23.
5. Richman DD, Wrin T, Little SJ, Petropoulos CJ. 2003. Rapid evolution of the neutralizing antibody response to HIV type 1 infection. Proc. Natl. Acad. Sci. U. S. A. 100:4144–4149.
6. Bhatt S, Holmes EC, Pybus OG. 2011. The genomic rate of molecular adaptation of the human influenza A virus. Mol. Biol. Evol. 28:2443–2451.
7. Hurst LD. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet. 18:486–487.
8. Chen L, Perlina A, Lee CJ. 2004. Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. J. Virol. 78:3722–3732.
9. Fernandes J, Jayaraman B, Frankel A. 2012. The HIV-1 rev response element: an RNA scaffold that directs the cooperative assembly of a homo-oligomeric ribonucleoprotein complex. RNA Biol. 9:4–9.
10. Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW, Jr, Swanstrom R, Burch CL, Weeks KM. 2009. Architecture and secondary structure of an entire HIV-1 RNA genome. Nature 460:711–716.
11. Keating CP, Hill MK, Hawkes DJ, Smyth RP, Isel C, Le S-Y, Palmenberg AC, Marshall JA, Marquet R, Nabel GJ, Mak J. 2009. The A-rich RNA sequences of HIV-1 pol are important for the synthesis of viral cDNA. Nucleic Acids Res. 37:945–956.
12. Forsdyke DR. 1995. Reciprocal relationship between stem-loop potential and substitution density in retroviral quasispecies under positive Darwinian selection. J. Mol. Evol. 41:1022–1037.
13. Snoeck J, Fellay J, Bartha I, Douek DC, Telenti A. 2011. Mapping of

14. Pollom E, Dang KK, Potter EL, Gorelick RJ, Burch CL, Weeks KM, Swanstrom R. 2013. Comparison of SIV and HIV-1 genomic RNA structures reveals impact of sequence evolution on conserved and nonconserved structural motifs. PLoS Pathog. 9:e1003294. doi:10.1371/journal.ppat.1003294.
15. Neher RA, Leitner T. 2010. Recombination rate and selection strength in HIV intra-patient evolution. PLoS Comput. Biol. 6:e1000660. doi:10.1371/journal.pcbi.1000660.
16. Batorsky R, Kearney MF, Palmer SE, Maldarelli F, Rouzine IM, Coffin JM. 2011. Estimate of effective recombination rate and average selection coefficient for HIV in chronic infection. Proc. Natl. Acad. Sci. U. S. A. 108:5661–5666.
17. Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H, Gupta P, Rinaldo CR, Learn GH, He X, Huang X-L, Mullins JI. 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J. Virol. 73:10489–10502.
18. Liu Y, McNevin J, Cao J, Zhao H, Genowati I, Wong K, McLaughlin S, McSweyn MD, Diem K, Stevens CE, Maenza J, He H, Nickle DC, Shriner D, Holte SE, Collier AC, Corey L, McElrath JM, Mullins JI. 2006. Selection on the human immunodeficiency virus type 1 proteome following primary infection. J. Virol. 80:9519–9529.
19. Bunnik EM, Pisas L, Van Nuenen AC, Schuitemaker H. 2008. Autologous neutralizing humoral immunity and evolution of the viral envelope in the course of subtype B human immunodeficiency virus type 1 infection. J. Virol. 82:7932–7941.
20. Kuiken C, Leitner T, Hahn B, Mullins J, Wolinsky S, Foley B, Apetrei C, Mizrachi I, Rambaut A, Korber B. 2012. HIV sequence compendium. Los Alamos National Laboratory, Los Alamos, NM.
21. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.
22. Zanini F, Neher RA. 2012. FFPopSim: an efficient forward simulation package for the evolution of large populations. Bioinformatics 28:3332–3333.
23. Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD. 1996. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science 271:1582–1586.
24. Markowitz M, Louie M, Hurley A, Sun E, Di Mascio M, Perelson AS, Ho DD. 2003. A novel antiviral intervention results in more accurate assessment of human immunodeficiency virus type 1 replication dynamics and T-cell decay in vivo. J. Virol. 77:5037–5038.
25. Levy DN, Aldovandi GM, Kutsch O, Shaw GM. 2004. Dynamics of HIV-1 recombination in its natural target cells. Proc. Natl. Acad. Sci. U. S. A. 101:4204–4209.
26. Josefsson L, Palmer S, Faria NR, Lemey P, Casazza J, Ambrozak D, Kearney M, Shao W, Kottilil S, Sneller M, Mellors J, Coffin JM, Maldarelli F. 2013. Single cell analysis of lymph node tissue from HIV-1 infected patients reveals that the majority of CD4+ T-cells contain one HIV-1 DNA molecule. PLoS Pathog. 9:e1003432. doi:10.1371/journal.ppat.1003432.
27. Boltz VF, Ambrose Z, Kearney MF, Shao W, KewalRamani VN, Maldarelli F, Mellors JW, Coffin JM. 2012. Ultrasensitive allele-specific PCR reveals rare preexisting drug-resistant variants and a large replicating virus population in macaques infected with a simian immunodeficiency virus containing human immunodeficiency virus reverse transcriptase. J. Virol. 86:12525–12530.
28. Neher RA, Shraiman B. 2011. Genetic draft and quasi-neutrality in large facultatively sexual populations. Genetics 188:975–996.
29. Asquith B, Edwards CTT, Lipsitch M, McLean AR. 2006. Inefficient cytotoxic T lymphocyte-mediated killing of HIV-1-infected cells in vivo. PLoS Biol. 4:e90. doi:10.1371/journal.pbio.0040090.
30. Moore PL, Ranchobe N, Lambson BE, Gray ES, Cave E, Abrahams M-R, Bandawe G, Mlisana K, Abdool Karim SS, Williamson C, Morris L, CAPRISA 002 study, NIAID Center for HIV/AIDS Vaccine Immunology (CHAVI). 2009. Limited neutralizing antibody specificities drive neutralization escape in early HIV-1 subtype C infection. PLoS Pathog. 5:e1000598. doi:10.1371/journal.ppat.1000598.
31. McKay MD, Beckman RJ, Conover WJ. 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21:239–245.

32. **Coffin JM.** 1995. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. Science **267:**483–489.

33. **Sanjuan R, Borderia AV.** 2011. Interplay between RNA structure and protein evolution in HIV-1. Mol. Biol. Evol. **28:**1333–1338.

34. **Jenkins GM, Holmes EC.** 2003. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res. **92:**1–7.

35. **van der Kuyl AC, Berkhout B.** 2012. The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus. Retrovirology **9:**92.

36. **Williamson S.** 2003. Adaptation in the *env* gene of HIV-1 and evolutionary theories of disease progression. Mol. Biol. Evol. **20:**1318–1325.

37. **Wei X, Decker JM, Wang S, Hui H, Kappes JC, Wu X, Salazar-Gonzalez JF, Salazar MG, Kilby JM, Saag MS, Komarova NL, Nowak MA, Hahn BH, Kwong PD, Shaw GM.** 2003. Antibody neutralization and escape by HIV-1. Nature **422:**307–312.

38. **Smith JM, Haigh J.** 1974. The hitch-hiking effect of a favourable gene. Genet. Res. **23:**23–35.

39. **Gillespie JH.** 2000. Genetic drift in an infinite population. The pseudo-hitchhiking model. Genetics **155:**909–919.

40. **Bar KJ, Tsao C-Y, Iyer SS, Decker JM, Yang Y, Bonsignori M, Chen X, Hwang K-K, Montefiori DC, Liao H-X, Hraber P, Fischer W, Li H, Wang S, Sterrett S, Keele BF, Ganusov VV, Perelson AS, Korber BT, Georgiev I, McLellan JS, Pavlicek JW, Gao F, Haynes BF, Hahn BH, Kwong PD, Shaw GM.** 2012. Early low-titer neutralizing antibodies impede HIV-1 replication and select for virus escape. PLoS Pathog. **8:**e1002721. doi:10.1371/journal.ppat.1002721.

41. **Fischer W, Ganusov VV, Giorgi EE, Hraber PT, Keele BF, Leitner T, Han CS, Gleasner CD, Green L, Lo C-C, Nag A, Wallstrom TC, Wang S, McMichael AJ, Haynes BF, Hahn BH, Perelson AS, Borrow P, Shaw GM, Bhattacharya T, Korber BT.** 2010. Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultradeep sequencing. PLoS One **5:**e12303. doi:10.1371/journal.pone.0012303.

42. **Hedskog C, Mild M, Jernberg J, Sherwood E, Bratt G, Leitner T, Lundeberg J, Andersson B, Albert J.** 2010. Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing. PLoS One **5:**e11345. doi:10.1371/journal.pone.0011345.

43. **Knoepfel SA, Berkhout B.** 2013. On the role of four small hairpins in the HIV-1 RNA genome. RNA Biol. **10:**512–524.

44. **Mayrose I, Doron-Faigenboim A, Bacharach E, Pupko T.** 2007. Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates. Bioinformatics **23:**i319–i327.

45. **Ngandu NK, Scheffler K, Moore P, Woodman Z, Martin D, Seoighe C.** 2008. Extensive purifying selection acting on synonymous sites in HIV-1 group M sequences. Virol. J. **5:**160.

46. **Ferguson AL, Mann JK, Omarjee S, Ndung'u T, Walker BD, Chakraborty AK.** 2013. Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. Immunity **38:**606–617.

47. **Josefsson L, King MS, Makitalo B, Brannstrom J, Shao W, Maldarelli F, Kearney MF, Hu W-S, Chen J, Gaines H, Mellors JW, Albert J, Coffin JM, Palmer SE.** 2011. Majority of CD4+ T cells from peripheral blood of HIV-1 infected individuals contain only one HIV DNA molecule. Proc. Natl. Acad. Sci. U. S. A. **108:**11199–11204.

48. **Nijhuis M, Boucher CAB, Schipper P, Leitner T, Schuurman R, Albert J.** 1998. Stochastic processes strongly influence HIV-1 evolution during suboptimal protease-inhibitor therapy. Proc. Natl. Acad. Sci. U. S. A. **95:**14441–14446.