Deleterious synonymous mutations hitchhike to high frequency in HIV *env* evolution

Fabio Zanini and Richard A. Neher

(Dated: February 13, 2013)

Intrapatient HIV evolution is dominated by selection on the protein level in the arms race with the adaptive immune system. When killer T cells or 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 analyze longitudinal intrapatient data from the C2-V5 part of the envelope gene (*env*) and observe that synonymous derived alleles rarely fix even though they often reach high frequencies in the viral population. 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, hinting at a direct fitness effect of these stem-loop structures in the HIV RNA. Computational modeling indicates that these synonymous mutations have a (Malthusian) selection coefficient of the order of -0.002, and that they are brought up to high frequency by hitchhiking on neighboring beneficial nonsynonymous alleles. The patterns of fixation of nonsynonymous mutations suggest that escape mutations in C2-V5 are only transiently beneficial, either because the immune system is catching up or because of competition between equivalent escapes.

I. INTRODUCTION

HIV 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 killer T cells (CTLs) and neutralizing antibodies (ABs) (Rambaut *et al.*, 2004) and facilitated by the high mutation rate of HIV (Abram *et al.*, 2010; Mansky and Temin, 1995). When the host develops a CTL or AB response against a particular HIV epitope, mutations in the viral genome that reduce or prevent recognition of the epitope frequently emerge. Escape mutations in epitopes targeted by CTLs typically evolve during early infection and spread rapidly through the population (McMichael *et al.*, 2009). During chronic infection, the most rapidly evolving part of the HIV genome are the variable loops V1-V5 in the envelope protein gp120, which change to avoid recognition by neutralizing ABs. Mutations in *env*, the gene encoding gp120, spread through the population within a few months (see Figure 1B). Consistent with this time scale, it is found that serum from a particular time typically neutralizes virus extracted more than 3-6 month earlier (Richman *et al.*, 2003).

These escape mutations are selected for their effect on the amino acid sequence of the viral proteins. Conversely, synonymous mutations are commonly used as approximately neutral markers in studies of viral evolution. Neutral markers are very useful since their dynamics can be compared to that of putatively functional sites to detect purifying or directional selection (Bhatt *et al.*, 2011; Chen *et al.*, 2004; Hurst, 2002). In addition to maintaining protein function and avoiding the adaptive immune recognition, however, the HIV 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 since these processes often depend on RNA folding. For example, the HIV *rev* response element (RRE) in *env* enhances nuclear export of full length or partially spliced viral transcripts via a complex stem-loop RNA structure (Fernandes *et al.*, 2012). Another well studied case is the interaction between viral reverse transcriptase, viral ssRNA, and the host $\mathrm{tRNA}^{\mathrm{Lys3}}$: the latter is required for priming reverse transcription (RT) and is bound by a pseudoknotted RNA structure in the viral 5' untranslated region (Barat *et al.*, 1991; Paillart *et al.*, 2002).

Even in absence of important RNA structures, synonymous codons do not evolve completely neutrally. Some codons are favored over others in many species (Plotkin and Kudla, 2011). Recent studies have shown that genetically engineered HIV strains with altered codon usage can in some cases produce more viral protein, but in general replicate less efficiently (Keating *et al.*, 2009; Li *et al.*, 2012; Ngumbela *et al.*, 2008). Codon deoptimization has been suggested as attenuation strategy for polio and influenza (Coleman *et al.*, 2008; Mueller *et al.*, 2010). Purifying selection beyond the protein sequence is therefore expected (Forsdyke, 1995; Snoeck *et al.*, 2011) and it has been shown that rates of evolution at synonymous sites vary along the HIV genome (Mayrose *et al.*, 2007). Positive selection through the host adaptive immune system, however, is restricted to changes in the amino acid sequence.

In this paper, we characterize the dynamics of synonymous mutations in *env* and show that a substantial fraction of these mutations is deleterious. We argue that synonymous mutations reach high frequencies via genetic hitchhiking due to the small recombination rate of HIV (Batorsky *et al.*, 2011; Neher and Leitner, 2010). We then compare our observations to computational models of HIV evolution and derive estimates for the effect synonymous mutations have on fitness. Extending the analysis of fixation probabilities to the nonsynonymous mutations, we show that time dependent selection or strong competition of escape mutations inside the same epitope are necessary to explain the observed patterns of fixation and loss.

II. RESULTS

The central quantity we investigate is the probability of fixation of a mutation, conditional on its population frequency. A neutral mutation segregating at frequency ν has a proba-

bility $P_{\rm fix}(\nu) = \nu$ to spread through the population and fix; in the rest of the cases, i.e. with probability $1-\nu$, it goes extinct. As illustrated in the inset of Fig. 1A, this is a simple consequence of the fact that (i) exactly one individual in the current population will be the common ancestor of the entire future population at a particular locus and (ii) this ancestor has a probability ν of carrying the mutation (assuming the neutral mutation is not preferentially associated with genomes of high or low fitness). Deleterious or beneficial mutations fix less or more often than neutral ones, respectively. Fig. 1 shows the time course the frequencies of all synonymous and nonsynonymous mutations observed *env*, C2-V5, in patient p10 (Shankarappa *et al.*, 1999), respectively. Despite many synonymous mutations reaching high frequency, few fix (panel 1A); in constrast, many nonsynonymous mutations fix (panel 1B).

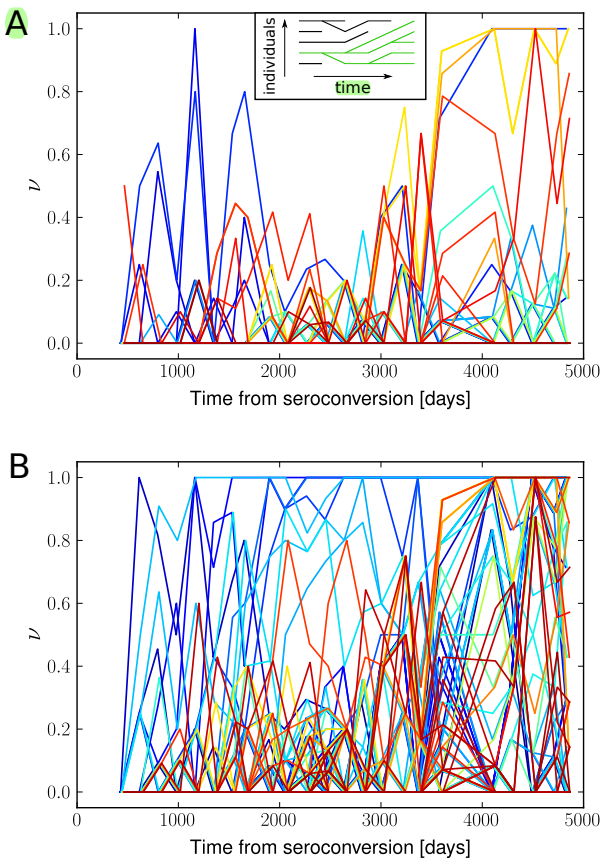

Figure 1 Time series of frequencies of synonymous (A) and nonsynonymous (B) derived alleles in *env*, C2-V5, from patient 10 (Shankarappa *et al.*, 1999). While many nonsynonymous mutations fix, few synonymous mutations do even though they are frequently observed at intermediate frequencies. Colors indicate the position of the site along the C2-V5 region (blue to red). Inset: the fixation probability $P_{\rm fix}$ of a neutral mutation is simply the likelihood that the future common ancestor is currently carrying it, i.e., its frequency ν.

A. Synonymous polymorphisms in *env*, C2-V5, are mostly deleterious

We study the dynamics and fate of synonymous mutations more quantitatively by analyzing data from 7 patients from Liu *et al.* (2006); Shankarappa *et al.* (1999) and 3 patients from Bunnik *et al.* (2008) (patients whose viral population was structured were excluded from the analysis; see methods and Figure S1). The former data set from is restricted to the C2-V5 region of *env*, while the data from Bunnik *et al.* (2008) covers the majority of *env*. Considering all mutations in a frequency interval around ν_0 at some time t, we calculate the fraction that is found at frequency 1, at frequency 0, or at intermediate frequency at later times $t+\Delta t$. Plotting these fixed, lost, and polymorphic fraction against the time interval Δt, we see that most synonymous mutations segregate for roughly one year and are lost much more frequently than expected (panel 2A). The long-time probability of fixation versus extinction of synonymous mutations is shown as a function of the initial frequency ν_0 in panel 2B (red line). Restricted to the region C2-V5, we find that $P_{\rm fix}$ of synonymous mutations is far below the neutral expectation. Outside of C2-V5, using data from Bunnik *et al.* (2008) only, no such reduction in $P_{\rm fix}$ is found. Restricted to the C2-V5 region, the sequence samples from Bunnik *et al.* (2008) are fully compatible with data from Shankarappa *et al.* (1999). The nonsynonymous mutations seem to follow more or less the neutral expectation (blue line) – a point to which we will come back below.

When interpreting these results for the fixation probabilities, it is important to distinguish between random mutations and polymorphisms observed at a certain frequency since the latter have already been filtered by selection. A polymorphism could be beneficial to the virus and on its way to fixation. In this case, we expect that it fixes almost surely given we see it at high frequency. If, on the other hand, the polymorphism is deleterious it must have reached a high frequency by chance (genetic drift or hitchhiking), and we expect that selection drives it out of the population again. Hence our observations suggest that many of the synonymous polymorphisms at intermediate frequencies in the part of *env* that includes the hypervariable regions are deleterious, while outside this regions most polymorphisms are roughly neutral. Note that this does not imply that all synonymous mutations in this region are neutral – only those mutations observed at high frequencies, which have been experiencing selection for some time already, tend to be neutral.

B. Synonymous mutations in C2-V5tend to disrupt conserved RNA stems

One possible explanation for lack of fixation of synonymous mutations in C2-V5are secondary structures in the viral RNA, the disruption of which is deleterious to the virus (Forsdyke, 1995; Sanjuan and Borderia, 2011; Snoeck *et al.*, 2011).

The propensity of nucleotides in the HIV genome to form base pairs has been measured using the SHAPE assay, a bio-

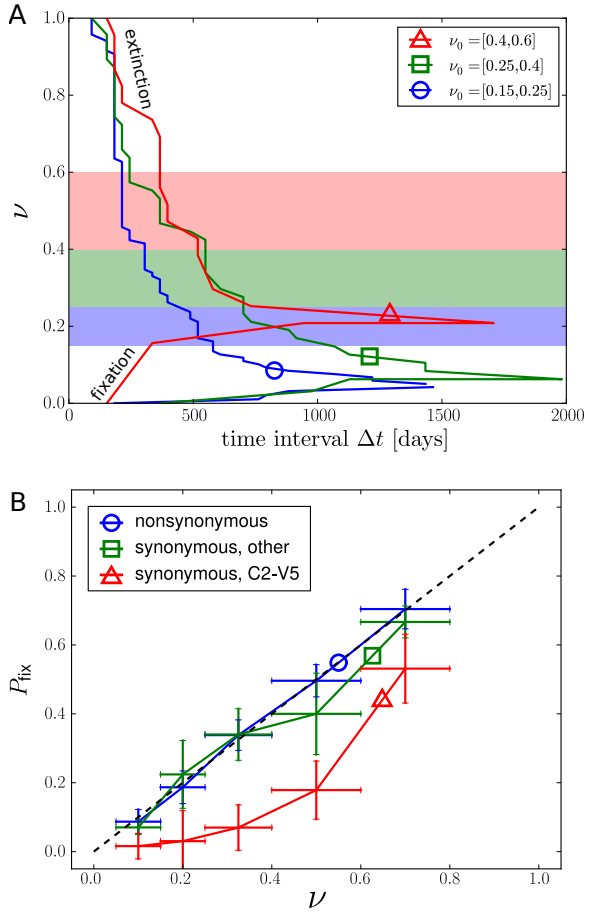

Figure 2 Fixation and loss of synonymous mutations. Panel A) shows the fraction of polymorphisms that have fixed (lower curves), are lost (distances between upper curves and 1), or remain polymorphic a time Δt later for different initial allele frequencies (colors, indicated as shaded area). In each frequency interval, the fraction of synonymous mutations that ultimately fix is the fixation probability conditional on the initial frequency. Panel B) shows the fixation probability of derived synonymous alleles as a function of ν_0. Polymorphisms within C2-V5 fix less often than expected for neutral mutations indicated by the diagonal line. This suppression is not observed in other parts of *env* or for nonsynonymous mutations. The horizontal error bars on the abscissa are bin sizes, the vertical ones the standard deviation after 100 patient bootstraps of the data. Data from refs. (Bunnik *et al.*, 2008; Liu *et al.*, 2006; Shankarappa *et al.*, 1999).

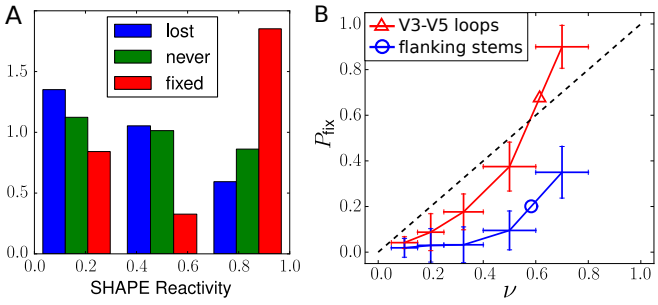

Figure 3 Permissible synonymous mutations tend to be unpaired. Panel A) shows the distribution of SHAPE reactivities among synonymous mutations that fix, mutations that reach frequencies 15%, but are susequently lost, and random mutations (all categ are restricted to the regions V1-V5±100bp). Fixed mutations tend to have higher higher SHAPE reactivities, corresponding to less base pairing. Mutations that are never observed show an intermediate distribution of SHAPE values. Panel B) shows the fixation probability of synonymous mutations in C2-V5 separately for variable loops and the connecting conserved regions that harbor RNA stems. As expected, the fixation probability is lower inside the conserved regions. Data from Refs. (Bunnik *et al.*, 2008; Liu *et al.*, 2006; Shankarappa *et al.*, 1999).

chemical reaction preferentially altering unpaired bases (the HIV genome is a single stranded RNA) (Watts *et al.*, 2009). The SHAPE assay has shown that the variable regions V1-V5 tend to be unpaired, while the conserved regions between those variable regions form stems. We partition all synonymous alleles observed at intermediate frequencies above 15% depending on their final destiny (fixation or extinction). Subsequently, we align our sequences to the reference NL4-3 strain used in ref. (Watts *et al.*, 2009) and assign them SHAPE reactivities. As shown in Fig. 3A, the reactivities of fixed alleles (red histogram) are systematically larger than of alleles that are lost (blue) (Kolmogorov-Smirnov test on the cumulative distribution, $p \approx 0.002$). In other words, alleles that are likely to break RNA helices are also more likely to revert and finally be lost from the population. The average over all mutations that are not observed (green) lies between the those that fix and those that get lost. Note that this analysis will be sensitive only at position where the base pairing pattern of NL4-3 agrees with that of each patient's initial consensus sequence (it is thus statistically conservative).

To test the hypothesis that mutations in C2-V5 are lost because they break stems in the conserved stretches between the variable loops, we consider mutations in variable loops and conserved parts separately. The biggest depression in fixation probability is observed in the conserved stems, while the variable loops show little deviation from the neutral signature, see Fig. 3B. This is consistent with important stem structures in conserved regions between loops.

In addition to RNA secondary structure, we have considered other possible explanations for a fitness defect of synonymous mutations, in particular codon usage bias (CUB). HIV is known to prefer A-rich codons over highly expressed human codons (Jenkins and Holmes, 2003; Kuyl and Berkhout, 2012). We do not find, however, any evidence for a contribution of average CUB to the ultimate fate of synonymous alleles; consistently, HIV does not seem to adapt its codon usage to its human host cells at the macroevolutionary level (Kuyl and Berkhout, 2012).

C. Deleterious mutations are brought to high frequency by hitchhiking

While the observation that some fraction of synonymous mutations is 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. This regions, however, undergoes frequent adaptive changes to evade recognition by neutralizing antibodies (Richman *et al.*, 2003; Williamson, 2003). Due to the limited amount of recombination in HIV (Batorsky *et al.*, 2011; Neher and Leitner, 2010), deleterious mutations that are linked to adaptive variants can reach high frequency. This process is known as hitchhiking (Smith and Haigh, 1974) or genetic draft (Gillespie, 2000; Neher and Shraiman, 2011). Hitchhiking is apparent in Fig. 1, that shows that many mutations change rapidly in frequency as a flock.

The approximate magnitude of the deleterious effects can be estimated from Fig. 2A, which shows the distribution of times for synonymous alleles to reach the fix or get lost starting from intermediate frequencies. The typical time to loss is of the order of 500 days. If this loss is driven by the deleterious effect of the mutation, this corresponds to deleterious effects of roughly $s_d \sim -0.002$ per day.

To get a better idea of the range of parameters that are compatible with the observations and our interpretation, we performed computer simulations 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 (Zanini and Neher, 2012). For each simulation run, we specify the deleterious effect of synonymous mutations, the fraction of synonymous mutations that are deleterious, the escape rate of adaptive nonsynonymous mutations and the frequency of new escapes. Note that the escape rate is the sum of two factors: (i) the beneficial effect due to the ability to evade the immune system minus (ii) the fitness cost of the mutation in terms of structure, stability, etc. Net escape rates in chronic infections have been estimated to be on the order of $\varepsilon = 0.01$ per day (Asquith *et al.*, 2006; Neher and Leitner, 2010).

Fig. 4A shows simulations results for the fixation probability and the synonymous diversity for different deleterious effects of synonymous mutations. We quantify synonymous diversity via P_{interm}, the fraction of sites with an allele at frequency $0.25 < \nu < 0.75$. The synonymous diversity observed in patient data is indicated in the figure. 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, lower inset). If no fixation happens, the area will be -0.5; if every mutation fixes, the area will be $+0.5$. In HIV infected patients, we find $A_{\text{syn}} \approx -0.2$ 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 alleles decreases from the neutral expectation ($A_{\text{syn}} \sim 0$) to zero ($A_{\text{syn}} \sim -0.5$) as their fitness effect worsens; the synonymous diversity plummets as well, as deleterious mutations

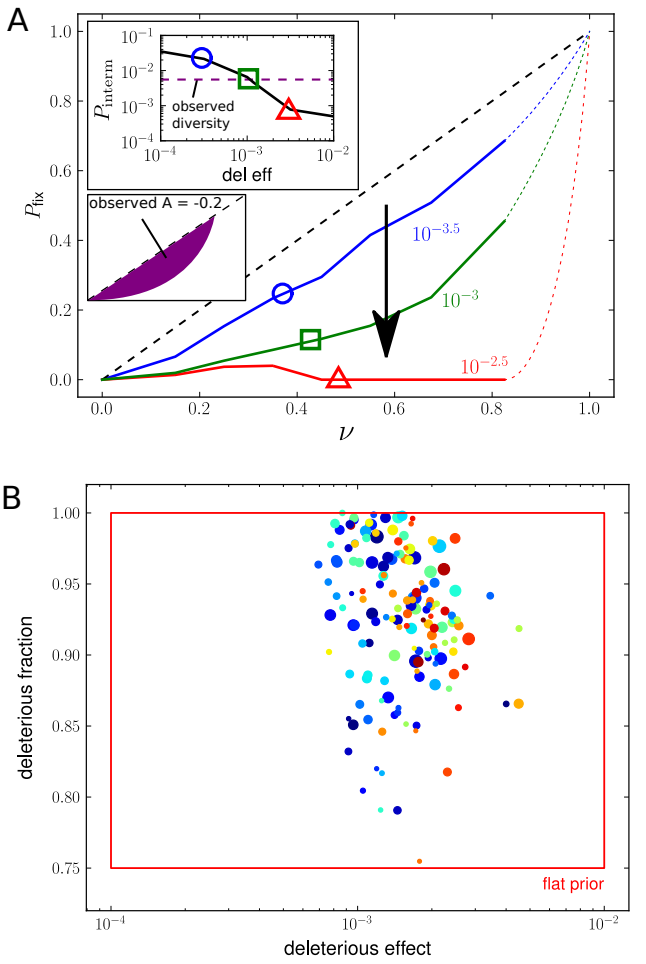

Figure 4 Distribution of selection coefficients on synonymous sites. Panel A) The depression in P_{fix} depends on the deleterious effect size of the synonymous alleles. This parameter also reduces synonymous diversity, measured by the probability of a derived allele to be found at intermediate frequencies P_{interm} (first inset). Panel B) To assess the parameter space that affects synonymous fixation and diversity, we run 2400 simulations with random parameters for deleterious effect size, fraction of deleterious synonymous sites, average size of escape mutations (color, blue ($10^{-2.5}$) to red ($10^{-1.5}$)), and rate of introduction of new epitopes (marker size, from 10^{-3} to 10^{-2} per generation). Mostly simulations with small deleterious effects and relatively high deleterious fractions of synonymous mutations reproduce synonymous the fixation probability and diversity observed in patients. Parameters are chosen from uniform prior distributions as indicated by the red box (see methods).

are selected against.

To map the parameter range of the model that is compatible with the data, we repeatedly simulated the evolution with random choices for the parameters in certain bounds, see Fig. 4B. Among all simulations, we select the ones that show A_{syn} and P_{interm} as observed in the data, i.e., a large depression in fixation probability of synonymous mutations but, simultaneously, a moderately high synonymous diversity. Specifically, Fig. 4B shows parameter combinations for which we found

$A_{\text{syn}} < -0.15$ and $0.0025 < P_{\text{interm}} < 0.010$. These conditions indicate that a high fraction ($\gtrsim 0.8$) of sites has to be deleterious with effect size $|s_d| \sim 0.002$. This result fits well the expectation based on the fixation/extinction times above (see Fig. 2A). The results are plausible: (i) a substantial depression in P_{fix} requires pervasive deleterious mutations, otherwise the majority of observed polymorphisms are neutral and no depression is observed; (ii) in order to hitchhike, the deleterious effect size has to be much smaller than the escape rate, otherwise the double mutant has little or no fitness advantage. Consistent with this argument, larger deleterious effects in Fig. 4 correspond to larger escapes rates. (iii) mutations with a deleterious effect smaller than approximately 0.001 behave neutrally consistent with the typical coalescent times observed in HIV.

The above simulation show that hitchhiking can explain the observation of deleterious mutations that rarely fix. However, in a simple model where nonsynonymous escape mutations are unconditionally beneficial, they almost always fix once they reach high frequencies – A_{nonsyn} is well above zero. This is incompatible with the blue line in Fig. 2: in an HIV infection, nonsynonymous mutations at high frequency often disappear again, even though many are at least transiently beneficial. Inspecting the trajectories of nonsynonymous mutations suggests the rapid rise and fall of many alleles. We test two possible mechanisms that are biologically sound and could explain the transient rise of nonsynonymous mutations: time-dependent selection and within-epitope competition.

The former hypothesis can be formulated as follows: 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, Bunnik *et al.* (2008); Richman *et al.* (2003) report antibody responses to 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. To model this type of behavior, we assume that antibody responses against escape mutations arise with a rate proportional to the frequency against the escape mutation and abolish the benefit of the escape mutations. As expected, this type of time dependent fixation retains the potential for hitchhiking, but reduces fixation of nonsynonymous mutations. Figure S3 shows that P_{fix} of synonymous mutations is not affected by the variance, while P_{fix} of nonsynonymous mutations approaches the diagonal as the rate of recognition of escape mutants is increased.

In the alternative hypothesis, several different escape mutations within the same epitope might arise almost simultaneously and start to spread. Their benefits are not additive, because each of them is essentially sufficient to escape. As a consequence, several escape mutations rise to high frequency rapidly, while the one with the smallest cost in terms of replication, packaging, etc. is most likely to eventually fix. The emergence of multiple sweeping nonsynonymous mutations in real HIV infections has been shown (Bar *et al.*, 2012; Moore *et al.*, 2009). Within epitope competition can be implemented the model through epistasis between escape mutations. While each mutation is individually beneficial, combining the mutations is deleterious (no extra benefit, but additional costs). Again, we find that the potential for hitchhiking is little affected by within epitope competition, but that the fixation probability of nonsynonymous polymorphisms is reduced. With roughly 6 mutations per epitope, the simulation data is compatible with observations; see Figure S4. The two scenarios are not exclusive and possibly both important in HIV evolution.

III. DISCUSSION

By analyzing the fate of mutations in longitudinal data of HIV *env* evolution, we demonstrate selection against synonymous substitutions in the comparatively conserved regions C2-C5 of the *env* gene. Comparison with biochemical studies of binding propensity of bases in RNA genome of HIV indicates that these mutations are deleterious, at least in part, because they disrupt stems in RNA secondary structures. Computational modeling shows that these mutations have deleterious effects on the order of 0.002 and that they are brought to high frequency through linkage to adaptive mutations.

The fixation and extinction times and probabilities represent a rich and simple summary statistics useful to characterize longitudinal sequence data and compare to models via computer simulations. A method that is similar to ours *in spiritu* has been recently used in a longitudinal study of influenza evolution (Strelkowa and Laessig, 2012). The propagators suggested in that article, however, represent ratios between nonsynonymous and synonymous mutations. The latter is used as an approximately neutral control; this method can therefore not be used to investigate synonymous changes themselves. More generally, evolutionary rates at synonymous sites are often used as a baseline to detect purifying or diversifying selection at the protein level (Hurst, 2002). It has been pointed out, however, that the rate of evolution at synonymous sites varies considerable along the HIV genome (Mayrose *et al.*, 2007) and that this variation can confound estimates of selection on proteins substantially (Ngandu *et al.*, 2008).

A functional significance of the insulating RNA structure stems between the hypervariable loops has also been proposed previously (Sanjuan and Borderia, 2011; Watts *et al.*, 2009) and conserved RNA structures exist in different parts of the HIV genome. Our analysis is able to quantify the fitness effect of RNA structure within single infections and demonstrates how selection at synonymous sites can alter genetic diversity and dynamics. The observed hitchhiking highlights the importance of linkage due to infrequent recombination for the evolution of HIV (Batorsky *et al.*, 2011; Josefsson *et al.*, 2011; Neher and Leitner, 2010). The recombination rate has been estimated to be on the order of $\rho = 10^{-5}$ per base and day. It takes roughly $t_{sw} = \varepsilon^{-1} \log \nu_0$ generations for escape mutation with escape rate ε to rise from an initially low frequency $\nu_0 \sim \mu$ to frequency one. This implies that a region of length $l = (\rho t_{sw})^{-1} = \varepsilon/\rho \log \nu_0$ remains linked to the adaptive mutation. With $\varepsilon = 0.01$, 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 regions, consistent with the lack of signal outside of C2-V5. In case of much stronger selection – such as observed during early CTL escape or drug resistance evolution – the linked region is of course much larger (Nijhuis *et al.*, 1998).

While classical population genetics assumes that the dominant stochastic force is genetic drift, i.e. non-heritable fluctuations in offspring number, our results show that stochasticity due to linked selection is much more important. Such fluctuations have been termed *genetic draft* by Gillespie (2000). Genetic draft in facultatively sexual population such as HIV has been characterized in (Neher and Shraiman, 2011). Importantly, large population sizes are compatible with low diversity and fast coalescence when draft dominates over drift.

Contrary to naïve expectations, the adaptive escape mutations do not seem to be unconditionally beneficial. Otherwise we would observe almost sure fixation of nonsynonymous mutations once they reach intermediate frequencies. Instead, we find that the fixation probability of nonsynonymous mutations is roughly given by its frequency. There are several possible explanations for this observation. Similar to synonymous mutations, the majority of nonsynonymous mutations could be weakly deleterious and the adaptive and deleterious parts conspire to yield a more neutral-like averaged fixation probability. We consider this possibility implausible since the amino acid sequence outside the variable loops is much more conserved that the synonymous positions, suggesting that the majority of the nonsynonymous mutations is much more deleterious.

Alternatively, the lack of fixation could be due to time dependent environment through an immune system that is catching up, or competition between mutations that mediate escape within the same epitope. We explore both of these possibilities and find that both produce the desired effect. Furthermore, there is experimental evidence in support of both of these hypotheses. Serum from HIV infected individuals typically neutralizes the virus that dominated the population a few (3-6) month ago (Richman *et al.*, 2003). This suggests that escape mutations cease to be beneficial after a few months and might revert if they come with a fitness cost. Deep sequencing regions of *env* after antibody escape have revealed multiple escape mutations in the same epitope (Bar *et al.*, 2012; Moore *et al.*, 2009). Presumably, each one of these mutations is sufficient for escape but most combinations of them do not provide any additional benefit to the virus. Hence only one mutation will spread and the others will be driven out of the population although they transiently reach high frequencies. The rapid emergence of multiple escape mutations in the same epitope implies a large population size that explores all necessary point mutations rapidly. A similar point has been made recently by Boltz *et al.* in the context of preexisting drug resistance mutations (Boltz *et al.*, 2012).

Our results emphasize the inadequacy of independent site models of HIV 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 (Batorsky *et al.*, 2011; Neher and Leitner, 2010) could be substantial underestimates.

IV. METHODS

A. Sequence data collection

Longitudinal intrapatient viral RNA sequences were collected from published studies (Bunnik *et al.*, 2008; Liu *et al.*, 2006; Shankarappa *et al.*, 1999) and downloaded from the Los Alamos National Laboratory (LANL) HIV sequence database (Kuiken *et al.*, 2012). The samples from some patients show substantial population structure and were discarded (see Figure S); a total of 11 patients with 4-23 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, most of them between 6 and 10 months.

B. Sequence analysis

The sequences were translated and the resulting amino acid sequences aligned using Muscle (Edgar, 2004) to each other and the NL4-3 reference sequences separately for each patient. Within each patient, the consensus nucleotide sequence at the first time point was used to classify alleles as “ancestral” or “derived” at all sites. Sites that include large 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.

C. Fixation probability and secondary structure

For the estimate of times to fixation/extinction, polymorphisms were binned by frequency and the time to first reaching either 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 got extinct were classified as “floating”, with one exception: if they first reached high frequencies within 3 years of the last time point, it was assumed they had not had sufficient time to settle, so they were discarded.

The SHAPE scores quantifying the degree of base pairing of individuals sites in the HIV genome were downloaded from the journal website (Watts *et al.*, 2009). 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 ref. (Watts *et al.*, 2009). 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 (Kuiken *et al.*, 2012).

D. Computer simulations

Computer simulations were performed using FFPopSim (Zanini and Neher, 2012). 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, mutations rates and arbitrary fitness functions can be specified. We used a generation time of 1 day, an outcrossing rate of $r = 0.01$ per day (Batorsky *et al.*, 2011; Neher and Leitner, 2010), a mutation rate of $\mu = 10^{-5}$ (Abram *et al.*, 2010; Mansky and Temin, 1995) and simulated intrapatient evolution for 6000 days. For simplicity, all third position in a codon where deemed synonymous and assigned either a selection coefficient 0 with probability $1-\alpha$ or a deleterious effect s_d with probability α. First and second positions have strongly deleterious fitness effects 0.02. At at rate k_A, a random place in the genome is designated an epitope that can escape by one or several mutations with an exponentially distributed escape rate with mean ε. Both full-length HIV genomes and *env*-only simulations were performed and yielded comparable results.

The simulations were repeated 2400 times with random choices for the following parameters: the fraction of deleterious sites α was sampled uniformly between 0.75 and 1.0; the average deleterious effect s_d was sampled such that its logarithm is uniformly distributed between 10^{-4} and 10^{-2}; the average escape rate ε of escape mutation such that its logarithm is uniform between $10^{-2.5}$ and $10^{-1.5}$; the rate k_A of new antibody challenges such that its logarithm is uniform between 10^{-3} and 10^{-2} per generation. Populations were initialized with a homogenous founder population and were kept at an average size of $N = 10^4$ throughout the simulation. After 30 generations of burn-in to create genetic diversity, new epitopes were introduced at a constant rate k_A.

For the models with competition within epitopes, a complex epistatic fitness landscape was designed such that each single mutant is sufficient for full escape. In particular, each mutation had a linear effect equal to the escape, but a negative epistatic effect of the same magnitude between each pair of sites was included. Higher order terms compensated each other to make sure that not only double, but all k-mutants with $k \geq 1$ had the same fitness (see supplementary materials). To model recognition of escape variants by the immune system that is catching up, 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.

For each set of parameters, fixation probabilities and probabilities of synonymous polymorphisms P_{interm} were calculated as averages over 100 repetitions (with different random seeds).

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. This measure is sufficiently precise for our purposes. In 10 runs out of 2400, the highest frequency bin was empty so the fixation probability could not be calculated; those runs were excluded from Fig. 4B.

E. Methods availability

All analysis and computer simulation scripts, as well as the sequence alignments used, are available for download.

Acknowledgements

We are grateful for stimulating discussions with Jan Albert and Trevor Bedford. This work is supported by the ERC starting grant HIVEVO 260686.

References

Abram, M. E., Ferris, A. L., Shao, W., Alvord, W. G., and Hughes, S. H. (2010). Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. *Journal of Virology*, **84**(19), 9864–9878.

Asquith, B., Edwards, C. T. T., Lipsitch, M., and McLean, A. R. (2006). Inefficient cytotoxic t lymphocyte-mediated killing of HIV-1-infected cells in vivo. *PLoS Biol*, **4**(4), e90.

Bar, K. J., Tsao, C.-y., Iyer, S. S., Decker, J. M., Yang, Y., Bonsignori, M., Chen, X., Hwang, K.-K., Montefiori, D. C., Liao, H.-X., Hraber, P., Fischer, W., Li, H., Wang, S., Sterrett, S., Keele, B. F., Ganusov, V. V., Perelson, A. S., Korber, B. T., Georgiev, I., McLellan, J. S., Pavlicek, J. W., Gao, F., Haynes, B. F., Hahn, B. H., Kwong, P. D., and Shaw, G. M. (2012). Early low-titer neutralizing antibodies impede HIV-1 replication and select for virus escape. *PLoS Pathog*, **8**(5), e1002721.

Barat, C., Grice, S. F. J. L., and Daelix, J.-L. (1991). Interaction of HIV-1 reverse transcriptase with a synthetic form of its replication primer, tRNALys,3. *Nucleic Acids Research*, **19**(4), 751–757.

Batorsky, R., Kearney, M. F., Palmer, S. E., Maldarelli, F., Rouzine, I. M., and Coffin, J. M. (2011). Estimate of effective recombination rate and average selection coefficient for HIV in chronic infection. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(14), 5661–6.

Bhatt, S., Holmes, E. C., and Pybus, O. G. (2011). The genomic rate of molecular adaptation of the human influenza a virus. *Molecular Biology and Evolution*, **28**(9), 2443–2451.

Boltz, V. F., Ambrose, Z., Kearney, M. F., Shao, W., KewalRamani, V. N., Maldarelli, F., Mellors, J. W., and Coffin, J. M. (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. *Journal of Virology*, **86**(23), 12525–12530.

Bunnik, E., Pisas, L., Van Nuenen, A., and 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. *Journal of virology*, **82**(16), 7932.

Chen, L., Perlina, A., and Lee, C. J. (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**(7), 3722–32.

Coleman, J. R., Papamichail, D., Skiena, S., Futcher, B., Wimmer, E., and Mueller, S. (2008). Virus attenuation by genome-scale changes in codon pair bias. *Science*, **320**(5884), 1784–1787.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**(5), 1792–1797.

Fernandes, J., Jayaraman, B., and Frankel, A. (2012). The HIV-1 rev response element: An RNA scaffold that directs the cooperative assembly of a homo-oligomeric ribonucleoprotein complex. *RNA Biology*, **9**(1), 4–9.

Forsdyke, D. (1995). Reciprocal relationship between stem-loop potential and substitution density in retroviral quasispecies under positive Darwinian selection. *Journal of Molecular Evolution*, **41**(6).

Gillespie, J. H. (2000). Genetic drift in an infinite population. the pseudohitchhiking model. *Genetics*, **155**(2), 909–19.

Hurst, L. D. (2002). The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet*, **18**(9), 486.

Jenkins, G. M. and Holmes, E. C. (2003). The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Research*, **92**(1), 1–7.

Josefsson, L., King, M. S., Makitalo, B., Brännström, J., Shao, W., Maldarelli, F., Kearney, M. F., Hu, W.-S., Chen, J., Gaines, H., Mellors, J. W., Albert, J., Coffin, J. M., and Palmer, S. E. (2011). Majority of CD4+ t cells from peripheral blood of HIV-1 infected individuals contain only one HIV DNA molecule. *Proceedings of the National Academy of Sciences*, **108**(27), 11199–11204.

Keating, C. P., Hill, M. K., Hawkes, D. J., Smyth, R. P., Isel, C., Le, S.-Y., Palmenberg, A. C., Marshall, J. A., Marquet, R., Nabel, G. J., and Mak, J. (2009). The A-rich RNA sequences of HIV-1 pol are important for the synthesis of viral cDNA. *Nucleic Acids Research*, **37**(3), 945–956.

Kuiken, C., Leitner, T., Hahn, B., Mullins, J., Wolinsky, S., Foley, B., Apetrei, C., Mizrachi, I., Rambaut, A., and Korber, B. (2012). *HIV Sequence Compendium 2012*. Theoretical Biology and Biophysics Group T-6, Mail Stop K710 Los Alamos National Laboratory Los Alamos, New Mexico 87545 U.S.A.

Kuyl, A. C. v. d. and Berkhout, B. (2012). The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus. *Retrovirology*, **9**(1), 92.

Li, M., Kao, E., Gao, X., Sandig, H., Limmer, K., Pavon-Eternod, M., Jones, T. E., Landry, S., Pan, T., Weitzman, M. D., and David, M. (2012). Codon-usage-based inhibition of HIV protein synthesis by human schlafen 11. *Nature*.

Liu, Y., McNevin, J., Cao, J., Zhao, H., Genowati, I., Wong, K., McLaughlin, S., McSweyn, M., Diem, K., Stevens, C., *et al.* (2006). Selection on the human immunodeficiency virus type 1 proteome following primary infection. *Journal of virology*, **80**(19), 9519.

Mansky, L. M. and Temin, H. M. (1995). Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *Journal of virology*, **69**(8), 5087–5094.

Mayrose, I., Doron-Faigenboim, A., Bacharach, E., and Pupko, T. (2007). Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates. *Bioinformatics*, **23**(13), i319–i327.

McMichael, A. J., Borrow, P., Tomaras, G. D., Goonetilleke, N., and Haynes, B. F. (2009). The immune response during acute HIV-1 infection: clues for vaccine development. *Nature Reviews Immunology*, **10**(1), 11–23.

Moore, P. L., Ranchobe, N., Lambson, B. E., Gray, E. S., Cave, E., Abrahams, M.-R., Bandawe, G., Mlisana, K., Abdool Karim, S. S., Williamson, C., Morris, L., the CAPRISA 002 study, and the 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**(9), e1000598.

Mueller, S., Coleman, J. R., Papamichail, D., Ward, C. B., Nimnual, A., Futcher, B., Skiena, S., and Wimmer, E. (2010). Live attenuated influenza virus vaccines by computer-aided rational design. *Nature Biotechnology*, **28**(7), 723–726.

Neher, R. and Leitner, T. (2010). Recombination rate and selection strength in HIV intra-patient evolution. *PLoS Comput Biol*, **6**(1), e1000660.

Neher, R. A. and Shraiman, B. (2011). Genetic draft and quasi-neutrality in large facultatively sexual populations. *Genetics*, **188**(4), 975–996.

Ngandu, N. K., Scheffler, K., Moore, P., Woodman, Z., Martin, D., and Seoighe, C. (2008). Extensive purifying selection acting on synonymous sites in HIV-1 group m sequences. *Virology Journal*, **5**(1), 160. PMID: 19105834.

Ngumbela, K. C., Ryan, K. P., Sivamurthy, R., Brockman, M. A., Gandhi, R. T., Bhardwaj, N., and Kavanagh, D. G. (2008). Quantitative effect of suboptimal codon usage on translational efficiency of mRNA encoding HIV-1 gag in intact T cells. *PLoS ONE*, **3**(6), e2356.

Nijhuis, M., Boucher, C. A. B., Schipper, P., Leitner, T., Schuurman, R., and Albert, J. (1998). Stochastic processes strongly influence HIV-1 evolution during suboptimal protease-inhibitor therapy. *Proceedings of the National Academy of Sciences*, **95**(24), 14441–14446.

Paillart, J.-C., Skripkin, E., Ehresmann, B., Ehresmann, C., and Marquet, R. (2002). In vitro evidence for a long range pseudoknot in the 5′-untranslated and matrix coding regions of HIV-1 genomic RNA. *Journal of Biological Chemistry*, **277**(8), 5995–6004.

Plotkin, J. B. and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics*, **12**(1), 32–42.

Rambaut, A., Posada, D., Crandall, K. A., and Holmes, E. C. (2004). The causes and consequences of HIV evolution. *Nature Reviews Genetics*, **5**(1), 52–61.

Richman, D. D., Wrin, T., Little, S. J., and Petropoulos, C. J. (2003). Rapid evolution of the neutralizing antibody response to HIV type 1 infection. *Proceedings of the National Academy of Sciences*, **100**(7), 4144–4149.

Sanjuan, R. and Borderia, A. V. (2011). Interplay between RNA structure and protein evolution in HIV-1. *Molecular Biology and Evolution*, **28**(4), 1333–1338.

Shankarappa, R., Margolick, J., Gange, S., Rodrigo, A., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C., Learn, G., He, X., *et al.* (1999). Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *Journal of Virology*, **73**(12), 10489.

Smith, J. M. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical research*, **23**(1), 23–35. PMID: 4407212.

Snoeck, J., Fellay, J., Bartha, I., Douek, D. C., and Telenti, A. (2011). Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural constraints. *Retrovirology*, **8**(1), 87.

Strelkowa, N. and Laessig, M. (2012). Clonal interference in the evolution of influenza. *Genetics*.

Watts, J. M., Dang, K. K., Gorelick, R. J., Leonard, C. W., Jr, J. W. B., Swanstrom, R., Burch, C. L., and Weeks, K. M. (2009). Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, **460**(7256), 711–716.

Williamson, S. (2003). Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression. *Molecular biology and evolution*, **20**(8), 1318–25.

Zanini, F. and Neher, R. A. (2012). FFPopSim: an efficient forward simulation package for the evolution of large populations. *Bioinformatics*.