Quantifying selection against synonymous mutations in HIV-1 env evolution

Fabio Zanini and Richard A. Neher
Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany

(Dated: August 18, 2013)

Contents

I. Selection of the patient data 2
II. Synonymous diversity across the HIV genome 3
III. Time-dependent selection 4
IV. Within-epitope competition 5
V. Marginal distributions in simulations 6

References 6
I. SELECTION OF THE PATIENT DATA

Figure S1 Structure of viral populations and patient selection. Panel A) shows a PCA of all sequences from patient p1 (colors indicate time from seroconversion, from blue to red). Panel B) shows allele frequency trajectories for nonsynonymous changes in the same patient. Here, the blue to red color map corresponds to the position of the allele in env from 5' to 3'. Panels C) and D) show analogous plots for data from patient p7. Samples after day 1000 split into two clusters in the PCA and no single nucleotide variants (SNVs) that arise after day 1000 fix, presumably because they are restricted to one subpopulation. All patients like p7 (p4, p7, p8, p9 from ref. Shankarappa et al., 1999 and ACH19542 and ACH19768 from ref. Bunnik et al., 2008) were excluded from our analysis.
II. SYNONYMOUS DIVERSITY ACROSS THE HIV GENOME

Figure S2 Synonymous diversity across the HIV genome, as quantified by the normalized codon entropy among sequences coding for the consensus amino acid. In most parts of the genome, synonymous sites show little conservation. The synonymous diversity peaks at the variable regions in env and is reduced in regions under purifying selection (RRE hairpin, second tat/rev exons). The normalized codon entropy is calculated as follows (see the script codon_entropy_synonymous_subtypeB.py for the full algorithm): (i) from a subtype B multiple sequence alignment (MSA) from the LANL website (filtered sequences only, version 2011) (Kuiken et al., 2012), we calculate the consensus amino acid at each position in the HIV genome; (ii) we count how often each codon coding for the consensus amino acid appears in the MSA; (iii) at each amino acid position, we divide by the number of sequences in the MSA that had the consensus amino acid at that position, obtaining codon frequencies $\nu_c$; (iv) we calculate the codon entropy from each position as: $S := -\sum_c \nu_c \log \nu_c$, where $c$ runs over codons that code for the consensus amino acid at this site; (v) we divide by the maximal codon entropy of that amino acid (e.g. $\log_2$ for twofold degenerate codons). All parts of env that are part of a different gene (signaling peptide, second rev exon) have been excluded from our main analysis, to avoid contamination by protein selection in a different reading frame. Note that all gap-rich columns of the MSA are stripped from this figure, so genes such as env might appear shorter than they actually are.
III. TIME-DEPENDENT SELECTION

Figure S3  Time-dependent selection reduces fixation of nonsynonymous SNVs. The figure compares the fixation probability in the time independent model (naïve) to a model with time dependent selection that mimics an evolving immune system. It has been found that virus is typically neutralized by serum from a few months earlier (Richman et al., 2003) but not by contemporary serum. We model this evolving immune system by assuming that escaped variants lose their beneficial effect with a rate proportional to the frequency of the escaped variant. Specifically, the selection effect of the escape mutations is reset to its fitness cost of $-0.02$ with probability $P_{\text{recognized}}(t) = c \cdot \nu(t)$, per generation, where $c$ is a constant coefficient shown in the legend that encodes the overall efficiency of the host immune system. With increasing probability of recognition, the fixation of frequent escape mutants is reduced, while hitch-hiking of synonymous SNVs is not affected. The precise shape of $P_{\text{fix}}(\nu)$ depends on the details of the $P_{\text{recognized}}(t)$, and we do not think that the high $P_{\text{fix}}(\nu)$ for $\nu < 0.2$ is meaningful. The other parameters for the shown simulations are the following: deleterious effect $s_d = 0.002$, average escape rate $\epsilon = 0.032$ per day, fraction of deleterious synonymous mutations $\alpha = 0.986$, rate of new epitopes $k_A = 0.0014$ per day.
IV. WITHIN-EPITOPE COMPETITION

Figure S4  Competition between escape SNVs in the same epitope reduces fixation of nonsynonymous SNVs. The figure compares the fixation probability of models with one, three, or six mutually exclusive escape mutations within the same epitope. Within epitope competition results in reduced fixation probabilities of nonsynonymous changes, whereas the synonymous changes behave similarly in all cases. We assume that escape can happen at \( n \) sites out of 3 consecutive codons and vary \( n \). The fitness landscape of each epitope includes negative epistatic terms, so that the joint presence of more than one escape mutation is not any more beneficial for the virus than a single mutation. Specifically, each site has two alleles, \( \pm 1 \), where \(-1\) is the ancestral one and \(+1\) the derived one; the fitness coefficient of a \( k \)-tuple of sites within the epitope is \( f_k = (-1)^{k-1} \epsilon^{1-n} \eta_k \), where \( \eta_k \) is the escape rate of the epitope drawn from an exponential distribution with mean \( \epsilon \) and \( n \) is the number of competing escapes in the epitope. In this evolutionary scenario, many escape SNVs start to sweep on different backgrounds within the viral population, but eventually compete and only one of them fixes. The other parameters for the shown simulations are the following: deleterious effect \( s_d = 0.002 \), average escape rate \( \epsilon = 0.032 \) per day, fraction of deleterious synonymous mutations \( \alpha = 0.986 \), rate of new epitopes \( k_A = 0.0014 \) per day.
Marginal distributions of simulation parameters

Figure S5 Marginal distributions of simulation parameters that reproduce the synonymous fixation probability and diversity seen in the data. Each panel contains the prior distribution of parameters (prior), the distribution of parameters after imposing the lower or upper bounds on synonymous diversity (syn diversity $\gtrless$), the bounds on the fixation probability (Area of $P_{\text{fix}}$), and the all bounds simultaneously (gated). The distribution of population sizes, the recombination rates, the escape rate of new epitopes, and the mutation rate remain approximately flat. They only two directions parameters whose distribution is strongly constrained by the data is the cost of synonymous mutations and the fraction of synonymous mutations that are neutral. We note that the distribution of compatible parameters is most likely a complicated subspace of the high dimensional parameter space and interpretation of these marginal projections is difficult.

References


