doi:10.1038/nature24634

Simulation of intrahost evolution

Intrahost HIV evolution was simulated using a stochastic model of birth, death, and mutation of infected cells. Each simulation starts with a single cell infected with the origin genotype. At a birth event, an actively replicating cell gives rise to another such cell, which has a chance of experiencing mutation at a random site (mutation rate u per site, sequenced DNA genome length L, HKY model-like nucleotide transition matrix with transition/transversion ratio κ and equilibrium frequencies p_A, p_C, p_G, p_T; origin genotype shares these frequencies). An actively replicating cell is "copied" without mutation to one of the three states identified in **Fig. 1** (fast-decay, medium-decay, or stable), at rate m for each state. At a death event, an actively replicating cell (per-cell rate d), fast-decay cell (rate d_1), or medium-decay cell (rate d_2) is removed from the population. Since the timescales considered are relatively short, cells in the stable reservoir persist throughout the duration of the simulation. When drawing sequence samples, we accounted for the fact that the number of cells in each state vary by several orders of magnitude (while avoiding simulation of a great many cells) by weighting each compartment appropriately.

Birth rate depends on the basic reproductive ratio (R_0), carrying capacity (C), current number of actively replicating cells (x), and genetic fitness for sequence i (f_i). Prior to treatment, the birth rate for sequence i is $b_i = \frac{dR_0C(f_i/\phi)}{C+x(R_0-1)}$, where ϕ is the average fitness among actively replicating cells. Note that at carrying capacity ($x = C$), for a virus of average fitness ($f_i = \phi$), the birth rate times the burst size equals the death rate, producing equilibrium. After treatment starts (four months after the start of infection), birth rate is set to zero. If sequence i carries no lethal mutations, then its fitness is $f_i = \exp(n_p s)$, where n_p is the number of positively-selected sites mutated in sequence i

(relative to the origin) and *s* is the selective advantage; if the sequence carries any lethal mutations, then its fitness is zero. To account for the effect of selection outside the sequenced region (e.g., selection in *env* when *gag* is sequenced), both beneficial and lethal sites may be included in the genome outside the region of length *L*.

Parameter values used in all simulations: Basic reproductive ratio $R_0 = 6$ (peak period, first 20 days of infection), 1.4 (after peak). Death rates $d = 1$/day, $d_1 = 0.165$/day (half-life of 4.2 days), $d_2 = 0.023$/day (half-life of 30 days). Mutation rate $u = 3 \times 10^{-5}$ per site per generation. Sequenced genome length $L = 587$ bp. Equilibrium nucleotide frequencies $p_A = 0.4$, $p_C = 0.15$, $p_G = 0.2$, $p_T = 0.25$. Transition/transversion ratio $\kappa = 3$. Per-cell state transition rate $m = 0.1$/day, with sampling weights 1.82 (fast-decay cells), 4.62×10^{-3} (medium-decay cells), and 5.13×10^{-6} (stable cells) relative to the active cells; these values result in ratios consistent with those determined in **Fig. 1**. During peak period, carrying capacity *C* was 50-fold higher than post-peak. Probability that a site within the sequenced region is lethal when mutated: 70%. Simulations started with a single infected cell and were implemented using discrete steps of 0.25 days; multiple events could occur in a single step.

Other parameters took 50 different sets of values. At least 197 replicates of each parameter set were run. Out of all ~12,000 runs, 77% were deemed to have a level of viral genetic diversity and divergence at the start of treatment consistent with that reported in studies of untreated patients: We considered average pairwise diversity between 0.092% and 0.285% per site (using the range of four patients with single-founder virus reported in Fig. 6 of Kearney *et al.*[1], interpolated to 120 days post-infection) and average divergence of 0.009% to 0.578% per site (using the range of estimated evolutionary rates of six patients reported in Fig. 5 of Alizon & Frazer[2]) to be realistic in *gag* and *pol* at four months post-infection.

Results summarized for each of the 50 parameter sets are shown in **Supplementary Table 1**, and detailed results for all runs are shown in **Supplementary Table 2**. **Fig. 2** of the main text depicts Replicate 91 of Parameter Set 48. A run is judged to have a misleading appearance of forward evolution if it returns a significantly increasing result ($p < 0.05$) on these three tests:

1. Comparison of average genetic divergence, measured from the most common genotype observed at the start of treatment, between the start of treatment and six months of treatment. (Mann-Whitney U test)
2. Estimate of evolutionary rate computed from linear regression of these divergences over time, using sequences sampled at the start of treatment and at three and six months of treatment. (F-test)
3. Estimate of evolutionary rate, or clock-like phylogenetic signal, computed from root-to-tip regression in the maximum likelihood phylogeny on sequences sampled at the start of treatment and at three and six months of treatment. (F-test)

Time-structured phylogenies

Time-resolved phylogenies were estimated using BEAST v2.4.4 with a single-compartment version of the model used by Lorenzo-Redondo et al. We used an HKY substitution model with no rate variation, strict molecular clock, and constant size population. More complex models were run but did not substantially improve fit nor change the results. Markov chain Monte Carlo sampling was run for 10^8 time-steps and convergence was verified using Tracer v1.6.0. The maximum clade credibility tree was chosen from sampled trees for visualization, and posterior clade probabilities >0.5 were annotated. The estimated evolutionary rate was 2.6×10^{-4} substitutions per site per month (95% highest posterior density interval 1.0×10^{-4} to 4.3×10^{-4}).

Simulation of post-treatment decay based on sequences observed during primary infection

As proposed in the Main Text, genetic patterns observed *after* ART has started depend on evolutionary dynamics *before* ART, since pre-ART dynamics determine the relative seeding of cells of varying longevity. The results documented in **Fig. 2** and **Supplementary Tables 1 and 2** rely on the above model of primary infection, which makes assumptions regarding population growth, mutation, and selection. Yet the dynamics of HIV are complex, and a model is always an approximation. Here we describe our approach to sidestep this limitation by using pre-treatment sequence data.

Novitsky *et al.*[3] longitudinally sampled individuals in Botswana infected with HIV-1 subtype C. Of the cohort of 32, three study participants had samples collected at least three times within the first four months of infection (IDs B, H, and OU). On average there were 13 sequences per time point, available on Genbank as described in the original study. We used only sequences of the *gag* gene (positions 841 to 2,217 of HXB2; *pol* was not sequenced). Sites containing gaps or ambiguous nucleotide calls were ignored. Participant B had 7 sites at which at least two sequences contained nucleotides not present at the first time-point, potential sites under selection. Participants H and OU had 2 and 12 such sites, respectively.

We assumed that the population at each timepoint was reported accurately by the sample. If a sample $\boldsymbol{S}_k$ of sequences (some possibly redundant) were observed at time t_k, we assumed that the true population at t_k was composed of these sequences at equal prevalence and no others. We further assumed that the population remained unchanged until t_{k+1}, at which time it instantaneously changed to match sample $\boldsymbol{S}_{k+1}$, based on the sequences observed at that time. We assumed that the first sequence, sampled near the time of seroconversion in all patients, represented the infection origin.

Post-ART population dynamics and simulated sampling followed the model described previously. We assumed that before ART, all viral strains contribute cells to a longer-lived compartment *j* at rate m_j, with values given above. When ART starts, we assume all viral replication is halted, so that seeding stops. All compartments decay at a rate d_j both before and after ART. The expected proportion of the simulated sample observed at time τ post-ART that matches a particular sequence present between times t_k and t_{k+1} is then

$$\ell_k(\tau) \propto \frac{1}{N_k}\sum_j \frac{m_j}{d_j}\left(1 - e^{-d_j(t_{k+1}-t_k)}\right)e^{-d_j(\tau+t_{ART}-t_k)},$$

where N_k is the number of sequences sampled at t_k and t_{ART} is the time at which ART is started (assumed to be 4 months). Note that this formula implicitly assumes that viral load is constant throughout primary infection. Early in infection, viral load starts low, but then peaks briefly above the eventual set-point. As the early timepoints sampled likely included periods both above and below set-point viral load, we chose not to include an explicit viral peak dynamic.

Post-ART samples for each time point (0, 3, and 6 months post-ART) were constructed by random sampling of 50 pre-ART sequences with replacement, weighted by $\ell_k(\tau)$. 100 replicates were analyzed by the same tests of forward evolution previously described for the fully simulated data. Results of each replicate are given in **Supplementary Table 3**; a summary is provided in **Extended Data Table 1**.

References

1. Kearney, M. *et al*. Human immunodeficiency virus type 1 population genetics and adaptation in newly infected individuals. *J. Virol.* **83,** 2715–2727 (2009).
2. Alizon, S., Fraser, C. Within-host and between-host evolutionary rates across the HIV-1 genome. *Retrovirology* **10**(49), 1–10 (2013).
3. Novitsky, V., Wang, R., Rossenkhan, R., Moyo, S. & Essex, M. Intra-host evolutionary rates in HIV-1C env and gag during primary infection. *Infect. Genet. Evol.* **19,** 361–368 (2013).