Supporting Information for

A binary trait model reveals the fitness effects of HIV-1 escape from T cell responses

Yirui Gao and John P. Barton

John P. Barton
E-mail: jpbarton@pitt.edu

This PDF file includes:

Supporting Information Text

Evolutionary model with selection on binary traits. In our study, model the evolution of a population of N individuals subject to recombination, mutation, and natural selection, following the Wright-Fisher (WF) model. For simplicity, we first describe a model with only two alleles per site, wild-type (WT) or mutant (MT). Individuals possess a genetic sequence length of L, resulting in a total of $M = 2^L$ genotypes. We further assume that there exist Λ binary traits, which depend on the presence or absence of mutant alleles at specific sites, and which are also subject to selection. For clarity, we will use $i, j, \ldots$ indices to denote the different loci or sites in the sequence, $n, m, \ldots$ are used for trait indices, and $a, b, \ldots$ represent genotype indices. Finally, $t_1, t_2, \ldots, t_K$ are used as generation (time) indices. The first two sets of indices are presented as subscripts, the genotype indices as superscripts, and the generation indices are displayed within parentheses to indicate quantities that vary with time.

Let $n_a(t_k)$ represent the number of individuals with genotype a at generation t_k, with $z_a(t_k) = n_a(t_k)/N$ the frequency of genotype a at generation t_k. The vector $z(t_k) = (z_a(t_k), z_b(t_k), \ldots, z_M(t_k))$ describes the state of the population at generation t_k. In our model, the probability of recombination occurring per site per generation is denoted by r. After recombination, the mean frequency of genotype a at generation t_{k+1}, is

$$y_a(t_{k+1}) = (1-r)^{L-1} z_a(t_k) + (1-(1-r)^{L-1})\psi_a(t_k)\,. \tag{1}$$

The term $(1-r)^{L-1}$ gives the likelihood of an individual not experiencing recombination, and $\psi_a(t_k)$ denotes the probability that the random recombination of any two individuals within the population results in an offspring of genotype a. This includes scenarios where both parent and offspring share the same genotype a.

After recombination, the mean frequency of each genotype in the next generation $p_a(t_k)$ is shaped by selection and mutation,

$$p_a(z(t_k)) = \frac{y_a(t_k) f_a + \mu \sum\limits_{b|d(a,b)=1} [y_b(t_k) f_b - y_a(t_k) f_a]}{\sum_{b=1}^{M} y_b(t_k) f_b}\,. \tag{2}$$

In this case, all alleles and loci share an equal mutation rate μ. For simplicity, we assume that each sequence undergoes at most one mutation per generation, given the exceedingly low mutation rate μ. The notation $b|d(a,b) = 1$ indicates that genotypes a and b differ by just a single mutation.

Due to the highly specific nature of TCR binding to peptide-MHC-I complexes, most nonsynonymous mutations within an epitope are likely to significantly disrupt binding, thereby facilitating immune escape. For any given epitope, different mutation paths can lead to a similar outcome: loss of recognition by T cells. Therefore, T cell escape for each individual epitope can be effectively modeled as a binary trait. The fitness of genotype a, denoted as f_a, is given by

$$f_a = 1 + \sum_{i}^{L} s_i g_i^a + \sum_{n}^{\Lambda} s_n g_n^a\,. \tag{3}$$

Contributions to fitness come from the effects of individual alleles (quantified by *selection coefficients* s_i) and traits (*trait coefficients* s_n). For the former, the impact of mutations is cumulative. In other words, the effects of mutant alleles at different sites add together. In contrast, for any given epitope n, the presence of one or more nonsynonymous mutations results in g_n being assigned a value of 1, irrespective of the number of these mutations. However, we emphasize that the fitness effects of different trait terms are additive, so that the effects of escape in two *different* T cell epitopes will add.

Path integral likelihood. Under WF dynamics, the probability of observing genotype frequencies $z(t_{k+1})$ at generation t_{k+1}, given genotype frequencies of $z(t_k)$= $(z_a(t_k), z_b(t_k), \ldots, z_M(t_k))$ at generation t_k, is

$$P(z(t_{k+1})|z(t_k)) = N! \prod_{a=1}^{M} \frac{(p_a(t_{k+1}))^{N z_a(t_{k+1})}}{(N z_a(t_{k+1}))!}\,. \tag{4}$$

Consequently, the likelihood that the genotype frequency vector follows a specific evolutionary trajectory, or "path", $\boldsymbol{z} = (z(t_1), z(t_2), \cdots, z(t_K)$, is

$$P(\boldsymbol{z}|z(t_0)) = \prod_{k=0}^{K-1} P(z(t_{k+1})|z(t_k))\,, \tag{5}$$

conditioned upon the initial state $z(t_0)$. This expression is difficult to work with directly, so we use several approximations to make our analysis more tractable.

To simplify this expression, we will project dynamics to the level of individual mutant loci i and trait groups n, instead of genotypes a. Here, we use the term "trait group" to refer to the set of all loci that contribute to the same $CD8^+$ T cell epitope. We denote the frequency of mutant alleles at site i in the population as x_i, and the frequency of individuals with one

or more nonsynonymous mutant alleles in trait group n as x_n. Additionally, the frequency of paired mutant alleles x_{ij} is used to describe the correlation between different mutations. The formulas for these are as follows:

$$\begin{aligned} x_i &= \sum_a^M g_i^a z_a \,, \\ x_{ij} &= \sum_a^M g_i^a g_j^a z_a \,, \\ x_n &= \sum_a^M g_n^a z_a = \sum_a^M \left[1 - \prod_{i\in n}(1-g_i^a)\right] z_a \end{aligned} \tag{6}$$

Here the term g_i^a indicates whether genotype a contains a mutant allele at locus i, with the wild type (WT) having $g_i^a = 0$ and the mutant type (MT) having $g_i^a = 1$. Similarly, g_n^a specifies whether trait group n contains a mutant allele. If all loci in trait group n of genotype a are WT, then $g_n^a = 0$; however, if there is at least one mutant allele in the trait group, meaning that at least one locus within the trait group has $g_i^a = 1$, then $g_n^a = 1$.

It is worth noting that, in most cases, trait groups can be regarded as a special type of locus. Consequently, we utilize the new subscripts $\boldsymbol{i}, \boldsymbol{j}$ to represent generic loci — encompassing both individual loci and trait groups — to distinguish them from individual loci i, j. Taking the pair allele frequency as an example, $x_{\boldsymbol{ij}}$ will represent not only the correlation between different individual loci but also between an individual locus and a trait group, as well as between different trait groups.

Next we consider the dynamics of the mutant allele frequencies (and trait frequencies) in the diffusion limit(1). We assume that the population size $N \to \infty$ and that the selection coefficients, mutation rate, and recombination rate are all small ($\mathcal{O}(1/N)$). In this limit, applying methods from statistical physics, the probability of an evolutionary trajectory $\boldsymbol{x} = (x(t_1), x(t_2), \ldots, x(t_K))$ can be quantified using a path integral (see refs. (2–4) for more details on this approach)

$$P(\boldsymbol{x}|x(t_0)) \propto \exp\left[-\frac{N}{2}\boldsymbol{S}(\boldsymbol{x})\right], \tag{7}$$

$$\boldsymbol{S}(\boldsymbol{x}) = \sum_{k=0}^{K-1} \frac{1}{\Delta t_k} \sum_{\boldsymbol{i},\boldsymbol{j}=1}^{L+\Lambda} \left[\Delta x_{\boldsymbol{i}}(t_k) - \Delta t_k D_{\boldsymbol{i}}(t_k)\right] \left[C_{\boldsymbol{ij}}(x(t_k))\right]^{-1} \left[\Delta x_{\boldsymbol{j}}(t_k) - \Delta t_k D_{\boldsymbol{j}}(t_k)\right] . \tag{8}$$

Here $\Delta t_k = t_{k+1} - t_k$ and $\Delta x_{\boldsymbol{i}}(t_k) = x_{\boldsymbol{i}}(t_{k+1}) - x_{\boldsymbol{i}}(t_k)$. $\boldsymbol{S}(\boldsymbol{x}(t_k))$ is referred to as the action in physics. In this expression, trait terms are considered as special individual loci, thus the total length of frequencies is the length of generic loci, which is $L + \Lambda$ (binary case). In statistical physics, $D_{\boldsymbol{i}}(t_k)$ and $C_{\boldsymbol{ij}}(t_k)/N$ are commonly referred to as the drift vector and the diffusion matrix respectively, which will be discussed in the following section. To prevent confusion, here we note that the drift vector quantifies the effects of natural selection, mutation, and recombination, which affect the expected change in allele frequencies, and *not* genetic drift.

The drift vector and diffusion matrix. In this section, unless otherwise specified, the time parameter for all physical quantities is assumed to be t_k, and the corresponding time indices will be omitted. We first begin with the diffusion matrix $C_{\boldsymbol{ij}}/N$, which is the same for both allele and trait frequencies

$$\frac{C_{\boldsymbol{ij}}}{N} = \langle (x_{\boldsymbol{i}}(t_{k+1}) - x_{\boldsymbol{i}}(t_k))(x_{\boldsymbol{j}}(t_{k+1}) - x_{\boldsymbol{j}}(t_k)) \rangle \approx \frac{1}{N} \begin{cases} x_{\boldsymbol{i}}(1 - x_{\boldsymbol{i}}) & \boldsymbol{i} = \boldsymbol{j} \,, \\ x_{\boldsymbol{ij}} - x_{\boldsymbol{i}} x_{\boldsymbol{j}} & \boldsymbol{i} \neq \boldsymbol{j} \,. \end{cases} \tag{9}$$

The expectation in the first line is taken over the WF model dynamics, which yields the second line in the limit that N is large and terms of $\mathcal{O}(1/N^2)$ are omitted. Here $\boldsymbol{i}, \boldsymbol{j}$ can be used in both individual and trait terms. $x_{\boldsymbol{i}}$ is the mutant frequencies at generic loci $\boldsymbol{i}$, while $x_{\boldsymbol{ij}}$ is the frequency of individuals with mutations at both generic loci $\boldsymbol{i}$ and $\boldsymbol{j}$ (see Eq. (6)). The diffusion matrix quantifies the amount of "noise" in changes in allele/trait frequencies due to finite population size N, i.e., genetic drift.

The drift vector describes the expected change in mutant allele frequencies in time due to selection, mutation, and recombination. For the trait frequencies, this term is especially complex.

Using Eq. (1) and Eq. (2) and dropping infinitesimals smaller than $\mathcal{O}(1/N)$, we find

$$\begin{aligned} D_{\boldsymbol{i}} &= \langle x_{\boldsymbol{i}}(t_{k+1}) - x_{\boldsymbol{i}}(t_k) \rangle \\ &= \sum_a g_{\boldsymbol{i}}^a (p_a(z) - z_a) \\ &\approx \sum_a^M g_{\boldsymbol{i}}^a \left\{ \left[\sum_j^L g_j^a s_j z_a - \sum_b^M \sum_j^L g_j^b s_j z_a z_b\right] + \left[\sum_n^\Lambda g_n^a s_n z_a - \sum_b^M \sum_n^\Lambda g_n^b s_n z_a z_b\right] + \left[\mu \sum_{b|d(a,b)=1} (z_b - z_a)\right] + \left[r(L-1)(\psi_a - z_a)\right] \right\} . \end{aligned} \tag{10}$$

Here we have four terms, which describe the evolutionary forces of selection for individual mutant alleles, selection on binary traits, mutation, and recombination, respectively.

Contributions due to natural selection can be written in the same way for both mutant alleles i and traits n (which, as a reminder, we group under a general bold-faced index $\boldsymbol{i}$). Using Eq. (6) and the expression for pair allele/trait frequencies $x_{\boldsymbol{ij}} = \sum_a^M g_{\boldsymbol{i}}^a g_{\boldsymbol{j}}^a z_a$, these terms become

$$\sum_j^L (x_{\boldsymbol{i}j} - x_{\boldsymbol{i}} x_j) s_j + \sum_n^\Lambda (x_{\boldsymbol{i}n} - x_{\boldsymbol{i}} x_n) s_n = \sum_{\boldsymbol{j}}^{L+\Lambda} C_{\boldsymbol{ij}} s_{\boldsymbol{j}} \,. \tag{11}$$

The mutation and recombination terms are different for mutant alleles and for traits. Here, we start with the mutation term for mutant alleles, which can be simplified as

$$\begin{aligned}
\mu \sum_a^M g_i^a \sum_{b|d(a,b)=1} (z_b - z_a) =& \mu \sum_a^M \sum_{\substack{b|d(a,b)=1 \\ g_i^a = g_i^b}} (g_i^b z_b - g_i^a z_a) + \mu \sum_a^M \sum_{\substack{b|d(a,b)=1 \\ g_i^a \neq g_i^b}} ((1-g_i^b) z_b - g_i^a z_a) \\
=& 0 + \mu \sum_a^M ((1 - g_i^a z_a) - g_i^a z_a) \\
=& \mu(1 - 2x_i) \,.
\end{aligned} \tag{12}$$

For the first term, $b|d(a,b) = 1\,,\ g_i^a = g_i^b$ means that a and b have the same allele at locus i, but they differ by one mutation at some other locus. If we sum over every genotype, these terms make no contribution to the change in mutant allele frequency i because every pair (a,b) has one conjugate pair (b,a). The second term is not symmetrical, but every a only has one b that can satisfy the conditions $b|d(a,b) = 1\,,\ g_i^a \neq g_i^b$ (i.e., the genotype with a MT allele at site i flipped to WT, or vice versa). Thus, we can use a to replace b and its summation.

Physically, Eq. (12) expresses the total flux due to mutation, which is the probability of mutation that can change the allele at locus i from WT to MT (flux in) minus the probability from MT to WT (flux out). The probability of MT at locus i is the mutant allele frequency. Since the state for the allele is binary, the sum of MT frequency and WT frequency is 1. Thus, we can easily simplify the expression as

$$\mu \sum_a^M g_i^a \sum_{b|d(a,b)=1} (z_b - z_a) = \mu(1 - x_i) - \mu x_i = \mu(1 - 2x_i) \,. \tag{13}$$

Next, we will consider the recombination term for mutant alleles. Similar to the mutation term, this represents the recombination flux for locus i. However, the expected mutant allele frequency change due to recombination alone is always zero. This is because, for every case in which a sequence without a mutant allele recombines with a sequence that has a mutant allele such that the recombinant sequence has the mutant allele (thus increasing the mutant allele frequency), there is another case with the MT and WT sequences switched (thus decreasing mutant allele frequency by the same amount) which occurs with the same probability. Thus, this term vanishes by symmetry.

In total, then, the drift vector for mutant frequencies i is

$$D_i = \sum_{\boldsymbol{j}}^{L+\Lambda} C_{i\boldsymbol{j}} s_{\boldsymbol{j}} + \mu\,(1 - 2x_i) \,. \tag{14}$$

Now we consider the mutation and recombination terms for traits. These terms are more complex because the conditions that can change the trait are different from the ones for a single locus, and generally depend on the state of all alleles that contribute to the trait. For example, mutation at one site within a trait group n will change the allele from WT to MT (or vice versa), but it may not change the state for the trait as a whole if there are other mutations among the trait group.

Let us consider the mutation term for traits, which we can expand as

$$\mu \sum_a^M g_n^a \sum_{b|d(a,b)=1} (z_b - z_a) = \mu \sum_a^M \Big[\sum_{\substack{b|d(a,b)=1 \\ g_n^a = g_n^b}} (g_n^b z_b - g_n^a z_a) + \sum_{\substack{b|d(a,b)=1 \\ g_n^a \neq g_n^b}} ((1-g_n^b) z_b - g_n^a z_a) \Big] \,. \tag{15}$$

As in Eq. (12), the first term is also zero. However, the second mutation term is different. In this case, genotypes a and b that satisfy $b|d(a,b) = 1\,,\ g_n^a \neq g_n^b$ do not exhibit a one-to-one correspondence: if all alleles in the trait group for a are WT, for example, then there are many b can satisfy these conditions and the number of b is the length of the trait group n. Alternately, if genotype a contains more than one mutant allele in the trait group, then it cannot be changed to WT within a single mutation.

To address this issue, we introduce a new variable, denoted as y_n^i, which is the frequency of genotypes that contain only one mutation in trait group n, and the mutation is at locus i. This can be written as

$$y_n^i = \sum_a^M g_n^a g_i^a \prod_{j \neq i, j \in n} (1 - g_j^a) z_a \,. \tag{16}$$

With this new variable, we can rewrite the mutation term for traits. Physically, to change the trait value (starting from WT), every locus among the trait group needs to be considered, as any mutation in these loci can change the state of the trait group from WT to MT. Conversely, mutation can only change the trait value from MT to WT if it affects genotypes that have one mutation among the trait group (i.e., by converting the single mutant allele to WT). Thus,

$$\mu \sum_{a}^{M} g_n^a \sum_{b|d(a,b)=1} (z_b - z_a) = \mu \sum_{i \in n} ((1 - x_n) - y_n^i) \,. \tag{17}$$

Finally, we turn to the recombination term for traits. Here we start by building physical intuition for the result. In each generation of the evolutionary process, recombination occurs numerous times. For this issue, we need to focus solely on the recombination that can alter the state of a trait group. In particular, we must enumerate the probabilities of recombination events that change traits *and* for which the reverse process (i.e., switching the order of the recombination partner sequences, which has the same probability) does not change the trait in the opposite way, thus yielding no net expected change in the trait frequency.

We use the index k to represent a recombination breakpoint. To change a trait's value, k must be located after the first site in the trait group and before the last. Since we have assumed that the recombination probability per site per generation r is small, we will only consider cases where a single recombination breakpoint falls within the trait sites, but the analysis below could be generalized to allow for two or more breakpoints. If recombination occurs between two sequences, where one is WT for trait group n and the other is MT before and after a breakpoint k for the same trait, then all the recombinants will be MT for trait group n. For such sequences, recombination can change WT to MT for the trait, but not MT to WT. This is the only scenario in which recombination leads to a net *increase* in trait frequency.

In the opposite direction, consider recombination between two sequences, where one is WT for trait n before a breakpoint k and MT after, and the second sequence is MT before k and WT after. Thus, both sequences are MT for trait n, but through recombination a WT sequence can be produced. This is the only recombination process that leads to a net *decrease* in trait frequency.

Collecting these two terms together, we arrive at the following expression for the net expected change in trait frequency due to recombination,

$$\sum_{a}^{M} g_n^a r(L-1)(\psi_a - z_a) = \psi_{a\_in}^n - \psi_{a\_out}^n = r \sum_{k \in n} P_{W,W}^{k,n} P_{M,M}^{k,n} - r \sum_{k \in n} P_{W,M}^{k,n} P_{M,W}^{k,n} \,. \tag{18}$$

Here $P_{M,W}^{k,n}$ represents the frequency of sequences that have at least one mutant allele in the trait group n on or before site k, and all WT alleles in the same trait group after k. In total, then, the drift vector for traits is

$$D_n = \langle x_n(t_{k+1}) - x_n(t_k) \rangle = \sum_{j}^{L+\Lambda} C_{nj} s_j + \mu \sum_{i \in n} ((1 - x_n) - y_n^i) + r \left(\sum_{k \in n} P_{W,W}^{k,n} P_{M,M}^{k,n} - \sum_{k \in n} P_{W,M}^{k,n} P_{M,W}^{k,n} \right) . \tag{19}$$

Two-step model of HIV-1 recombination. As noted in the main text, HIV-1 recombination actually occurs in multiple steps, including the coinfection of a single cell by two distinct viruses and reverse transcriptase template switching between the two distinct RNA strands. Let us refer to the probability of coinfection as p_c and the probability of template switching as p_s per base per replication cycle. We explored how this two-step model of recombination would affect our estimates of selection.

As before, we must search for recombination events that could affect the escape frequency for a particular epitope. In this model as well, recombination breakpoints must occur within an epitope (i.e., after the first site in the epitope and before the last site) to affect escape frequency. Previously, we had assumed a single effective recombination rate per site $r \ll 1$. Now, we break this into two steps, with probability p_c for coinfection and p_s per base for template switching. Naively relating $r \sim p_c p_s$, the difference between these models becomes clear: the probability of observing two recombination breakpoints within some region is proportional to r^2 in the one-step model, but $p_c p_s^2$ in the two-step model.

Previously, we had assumed that multiple recombination events within the same epitope should be very unlikely, and that the dominant contribution to changes in escape frequency should therefore come from single recombination events (i.e., $r^2 \ll r$). To make the same approximation in the two-step recombination model, we would need to show that $p_c p_s^2 \ll p_c p_s$. While estimates for the template switching probability vary, high-end estimates are generally around $p_s \sim 10^{-3}$. There is also evidence for potential hotspots of recombination in HIV-1, where template switching may occur more frequently(5–8). However, even if template switching were to occur at a rate of 10^{-2} per base in a recombination hot spot, the assumption that $p_c p_s \ll p_c p_s^2$ is likely to be a very reasonable one. Thus, to leading order in the template switching rate, one could replace the more biologically realistic two-step recombination process with an effective one-step process with an effective rate $r = p_c p_s$ to extract the dominant contributions of recombination to escape frequency change (and therefore to inferred selection).

Marginal path likelihood (MPL) inference. To find the fitness effects of alleles/traits that best fit the data, we will attempt to find parameters s_i that maximize the likelihood of the data. Since the likelihood P is proportional to action $\boldsymbol{S}$ (see Eq. (7)), the unknown selection coefficients s_i and trait coefficients s_n that can maximize action $\boldsymbol{S}$ will maximize the path probability P.

To control our estimates, we also introduce a Gaussian prior distribution for the selection coefficients with mean zero and precision γN,

$$P_{prior}(\boldsymbol{s}(t)) = \frac{1}{(2\pi\sigma^2)^{(L+\Lambda)/2}} \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{s}^T\boldsymbol{s}\right) \propto \exp\left(-\frac{N}{2}\gamma\boldsymbol{s}^2\right). \tag{20}$$

Here we have absorbed the population size N into the width of the prior for convenience (see below). Including the prior distribution, the action becomes

$$\boldsymbol{S}(\boldsymbol{x}(t_k)) = \sum_{k=0}^{K-1}\frac{1}{\Delta t_k}\sum_{\boldsymbol{i},\boldsymbol{j}=1}^{L+\Lambda}\left\{[\Delta x_{\boldsymbol{i}}(t_k) - \Delta t_k D_{\boldsymbol{i}}(t_k)] \times [C_{\boldsymbol{ij}}(x(t_k))]^{-1}[\Delta x_{\boldsymbol{j}}(t_k) - \Delta t_k D_{\boldsymbol{j}}(t_k)]\right\} + \sum_{\boldsymbol{i}}^{L+\Lambda}\gamma s_{\boldsymbol{i}}^2 \tag{21}$$

Finally, we can apply Bayes' theorem to infer selection coefficients $\hat{s}_{\boldsymbol{i}}$ that maximize the posterior distribution, providing the best compromise between the prior and the likelihood. While the posterior distribution is a complicated function of allele/trait frequencies, it is simply a Gaussian function of the selection coefficients. This allows us to write an analytical expression for the maximum a posteriori parameters, as shown in Eq. (2) in the main text, with

$$\boldsymbol{C}_{int} = \sum_{k=0}^{K-1}\Delta t_k \boldsymbol{C}(t_k), \tag{22}$$

$$\Delta\boldsymbol{x} = \boldsymbol{x}(t_K) - \boldsymbol{x}(t_0), \tag{23}$$

$$\boldsymbol{\mu}_{\text{fl}} = \begin{cases} \sum_{k=0}^{K-1}\Delta t_k \mu(1 - 2\boldsymbol{x}(t_k)), \\ \sum_{k=0}^{K-1}\Delta t_k \mu \sum_{i\in N}(1 - \boldsymbol{x}_n(t_k) - \boldsymbol{y}_n^i(t_k)), \end{cases} \tag{24}$$

$$\boldsymbol{R}_{\text{fl}} = \begin{cases} 0, \\ \sum_{k=0}^{K-1}\Delta t_k r\left(\sum_{k\in n} P_{W,W}^{k,n}P_{M,M}^{k,n} - \sum_{k\in n} P_{M,W}^{k,n}P_{W,M}^{k,n}\right). \end{cases} \tag{25}$$

Here, the top expressions for $\boldsymbol{\mu}_{\text{fl}}$ and $\boldsymbol{R}_{\text{fl}}$ are for mutant alleles, and the bottom ones are for traits. In simulations, we used $\gamma = 1$ for selection coefficients and 0.1 for traits. In noisier HIV-1 data, we used $\gamma = 10$ for selection coefficients and 1 for traits.

Extension to multiple alleles per locus and asymmetric mutation probabilities. To study real sequences, we can extend the simple model presented in the previous sections to allow for multiple alleles per locus and asymmetric mutation probabilities. We use $\alpha, \beta, \ldots$ indices to represent different alleles (ranging, for example, over nucleotides or amino acids), where we write the total number of possible alleles at a locus as l. Our fitness model is then:

$$f_a = 1 + \sum_{i}^{L}\sum_{\alpha}^{l} s_{i,\alpha}g_{i,\alpha}^a + \sum_{n}^{\Lambda} s_n g_n^a. \tag{26}$$

Similarly, $s_{i,\alpha}$ represents the selection coefficient for allele α at locus i, and $g_{i,\alpha}^a$ equals 1 if genotype a has allele α at locus i. The trait term is the same as the binary case since it does not have a natural counterpart (only 2 states, WT and MT). We use $x_{i,\alpha}(t_k)$ to represent the frequency of allele α at locus i at generation t_k, and $\mu_{\alpha\beta}$ to denote the probability per locus per generation of mutation from allele α to β. Following parallel arguments to before, the MPL estimate of the selection coefficient $s_{i,\alpha}$ for each allele α at each locus i and the trait coefficients s_n can be obtained.

First, we write the diffusion matrix $C_{i\alpha,j\beta}(t_k)/N$,

$$C_{\boldsymbol{i}\alpha,\boldsymbol{j}\beta}(t_k) = x_{\boldsymbol{i}\alpha,\boldsymbol{j}\beta}(t_k) - x_{\boldsymbol{i}\alpha}(t_k)\cdot x_{\boldsymbol{j}\beta}(t_k). \tag{27}$$

where $x_{\boldsymbol{i}\alpha,\boldsymbol{j}\beta}(t_k)$ is the frequency of sequences with alleles α and β at loci $\boldsymbol{i}$ and $\boldsymbol{j}$ at generation t_k. When one of the indices corresponds to a trait, allele subscripts are not needed. For example, the covariance between trait group n and allele α at locus i can be written as $C_{n,i\alpha}(t_k) = x_{n,i\alpha}(t_k) - x_n(t_k)\cdot x_{i\alpha}(t_k)$.

The estimated selection coefficients for alleles are then

$$s_{i,\alpha} = \sum_{\boldsymbol{j}}^{L+\Lambda}\sum_{\beta}^{l}\left[\sum_{k=0}^{K-1}\Delta t_k C_{i\alpha,\boldsymbol{j}\beta}(t_k) + \gamma\hat{I}\right]^{-1}_{i\alpha\boldsymbol{j}\beta}\cdot\left[x_{\boldsymbol{j}\beta}(t_K) - x_{\boldsymbol{j}\beta}(t_0) - \sum_{k=0}^{K-1}\Delta t_k\sum_{\delta\neq\alpha}(\mu_{\delta\alpha}x_{\boldsymbol{j}\delta}(t_k) - \mu_{\alpha\delta}x_{\boldsymbol{j}\alpha}(t_k))\right], \tag{28}$$

and the selection coefficients for traits are given by

$$s_n = \sum_{\boldsymbol{j}}^{L+\Lambda} \sum_{\beta}^{l} \left[\sum_{k=0}^{K-1} \Delta t_k C_{n,\boldsymbol{j}\beta}(t_k) + \gamma \hat{I} \right]^{-1}_{n\boldsymbol{j}\beta} \cdot \Big[x_{\boldsymbol{j}\beta}(t_K) - x_{\boldsymbol{j}\beta}(t_0) - \sum_{k=0}^{K-1} \Delta t_k \sum_{i\in n} \sum_{\delta\neq\epsilon} \sum_{\epsilon} \left(\mu_{\epsilon\delta} y_n^{i\epsilon} - \mu_{\delta\epsilon} y_n^{i\delta}(t_k) \right) - \sum_{k=0}^{K-1} \Delta t_k r \left(\sum_{k\in n} P^{k,n}_{W,W} P^{k,n}_{M,M} - \sum_{k\in n} P^{k,n}_{M,W} P^{k,n}_{W,M} \right) \Big] . \quad [29]$$

Here the index ϵ runs over WT alleles and synonymous mutations for each binary trait n. This is because only nonsynonymous mutations contribute to changes in the trait in our model.

In this version of the model, we compute selection coefficients for both mutant and WT alleles. However, what is important is not the selection coefficient for an individual allele in isolation, but rather the change in fitness upon mutation. Thus, the selection coefficients we report are normalized by subtracting the WT selection coefficient from the selection coefficient of each allele at each locus ($s_{i\alpha} - s_{iW}$). In this way, the selection coefficients are normalized such that the WT selection coefficient is 0, and the selection coefficients for other alleles quantify the advantage or disadvantage relative to WT. In the limit that the selection coefficients are small, one can show that the WF dynamics are invariant under such shifts in the selection coefficients. In physics, this phenomenon is referred to as gauge invariance.

Simulation data. We simulated the WF model with discrete generations and binary (mutant/WT) states in Python. Briefly, we evolved populations of sequences according to Eq. (4) over multiple generations, starting with an initial population of all WT sequences and recording the entire evolutionary history. After simulation, we computed the single x_i, double x_{ij} mutant frequencies, and trait frequencies x_n from the sampled sequences trajectories and used them to infer individual selection coefficients (Eq. (28)) and trait coefficients (Eq. (29)). Parameter values are detailed in **Supplementary Fig. S1**. The simulation and analysis code with original simulation data are contained in the GitHub repository.

Since real data typically contains only a small portion of population, and is not sampled at every generation, we also studied how different sampling depths and sampling time intervals affect the performance of our method. We chose part of the sequences from the population and time points to estimate individual selection and trait coefficients. n_s denotes the number of sequences we randomly selected from the population and Δt is the time interval, which means we choose the data every Δt generations. The initial population and simulation parameters are described in **Supplementary Fig. S2**.

$CD8^+$ T cell levels can fluctuate over time, suggesting that a time-varying binary trait selection coefficient would most accurately describe this phenomenon. Thus, we also conducted a simulation with time-varying selection coefficients of binary traits to test our approach. The results in **Supplementary Fig. S1** demonstrate that the trait coefficients we estimate are approximately equivalent to the average trait coefficient when selection varies over time. The parameter values are detailed in **Supplementary Fig. S1**.

HIV-1 sequence data. We obtained HIV-1 sequence data from 13 individuals of the CHAVI 001 and CAPRISA 002 studies in the United States, Malawi, and South Africa from the Los Alamos National Laboratory (LANL) HIV Sequence Database. We applied several selection criteria(3) to minimize the influence of noise in the data, including removing the sequences with large numbers of gaps, sites with high gap frequencies, and time points with very small numbers of sequences or large gaps in time from the last sample. We also imputed ambiguous nucleotides with the most common nucleotides observed at the same site within the same individual.

For these 13 individuals, sequence data consisted of $3'$ and $5'$ half-genome sequences, which were approximately 4,500 bp in length. Our analysis focused only on polymorphic sites, where more than one nucleotide (including gaps/deletions) was observed in an individual (approximately 100-900 bp in length). To infer selection, we used a mutation rate matrix estimated in ref. (9) as input. We allowed the effective recombination rate r to vary along with viral load (VL), following recent work that revealed increasing effective recombination rates in individuals with higher VL due to higher levels of coinfection (10). To estimate r as a function of VL, we used a linear model with parameters roughly fit to the data of Romero and Feder ($r = 1.722 \times 10^{-10}\,\mathrm{VL} + 1.39 \times 10^{-5}$). In our model, we dynamically determined the recombination rate based on the viral load at each time point, which was measured in past work (11). For later stages of infection where VL was not measured, we assumed that VL values remained unchanged from the most recent measurement, consistent with the establishment of a viral set point in chronic infection. Although there are large spikes in VL during acute infection (resulting in a corresponding recombination rate on the order of 10^{-3}, compared to typical constant estimates of around 1.4×10^{-5} (ref. (12))), it quickly settles down to a value orders of magnitude lower. We treated the transmitted/founder (TF) sequence as the "wild-type" sequence for each individual.

The locations of $CD8^+$ T cell epitopes for these sequences were experimentally(11) or computationally(13) determined. In order to disentangle the fitness effects escape from the effects of individual mutations, we focus on escape effects that can be inferred independently from other contributions to fitness. This requires that the escape trait should be neither completely correlated nor anti-correlated with other variants. Mathematically, we obtain the selection and escape coefficients that best fit the data by solving a linear equation, Eq. (2), which has the form $Ax = b$. Here, the matrix $A = C_{\mathrm{int}} + \gamma I$ and the vectors $x = \hat{\boldsymbol{s}}$ and $b = \Delta \boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{fl}} - \boldsymbol{R}_{\mathrm{fl}}$. One can then compute the reduced row echelon form of the matrix A to determine which parameters are linearly independent(14). The regularization term γI renders all parameters trivially linearly independent, so to identify which parameters are linearly independent based on the data alone, we computed the reduced row echelon form of the matrix

C_{int} for each individual and sequencing region. The escape coefficients which are dependent on other variables correspond to the epitopes totally correlated with other mutations. These mutations could be either the individual escape mutations in the epitope or the mutation outside the epitope due to a lack of observed sequence data. We then selected only the escape coefficients that were linearly independent from all other variables for analysis (i.e., the "accessible" escape coefficients, in the language of ref. (14)). These typically encompass epitopes containing three or more loci with nonsynonymous mutations, though in some cases, two loci may suffice. *Escape sites* refer to polymorphic sites where nonsynonymous mutations were observed in the reading frame of an independent (accessible) $CD8^+$ T cell epitope. In this way, we anticipate that nonsynonymous mutations in escape sites will affect T cell recognition. We consider the escape sites that can change the same epitope to be part of a single "trait group."

Calculation of effects of linkage on inferred selection. $\Delta\hat{s}_{ij}$ can tell us the effects of linkage from variant i to j. To compute it, we calculate the coefficients for mutant variant j and eliminate the influence from mutant variant i by artificially reverting variant i to WT(3).

For mutant alleles, we generated a modified version of the sequence data where all mutant variants i are replaced by the corresponding TF nucleotide for all sequences at all time points. For traits, we treat all mutant variants within one epitope as WT. With these modified data, we can infer the coefficients again for all variants $\boldsymbol{j}$, denoted as $\hat{s}_j^{\setminus i}$. Then we define

$$\Delta\hat{s}_{ij} = \hat{s}_j - \hat{s}_j^{\setminus i}. \quad [30]$$

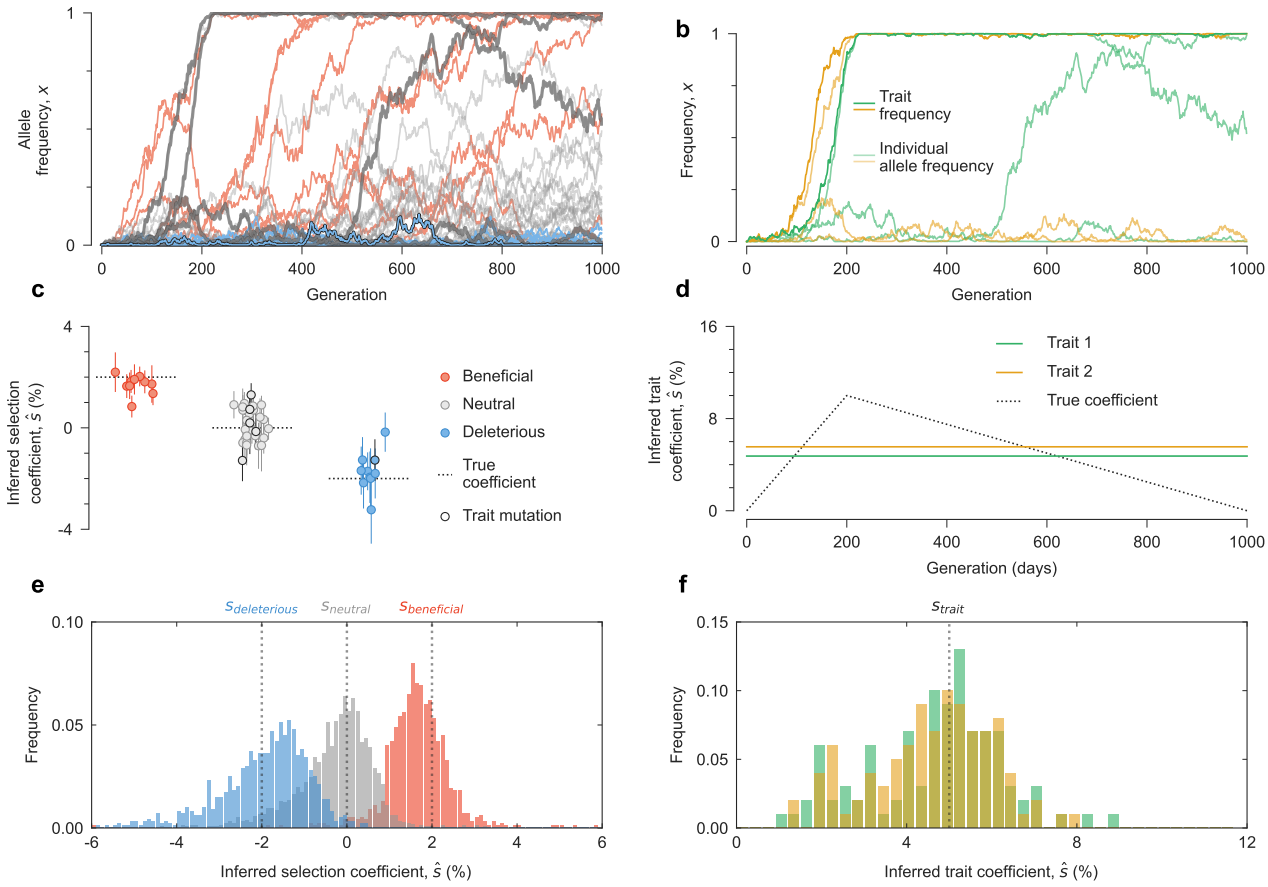
Positive values of $\Delta\hat{s}_{ij}$ indicate that linkage with variant i can increase the selection coefficient inferred for variant j and vice versa. By computing the $\Delta\hat{s}_{ij}$, we can quantify the effect of linkage on inferred selection.

The effects of recombination on the inference. Unlike selection on individual alleles, which has been studied previously using similar approaches(3, 4), the selection that we infer for binary traits is affected by recombination. How large is the contribution of recombination in this analysis?

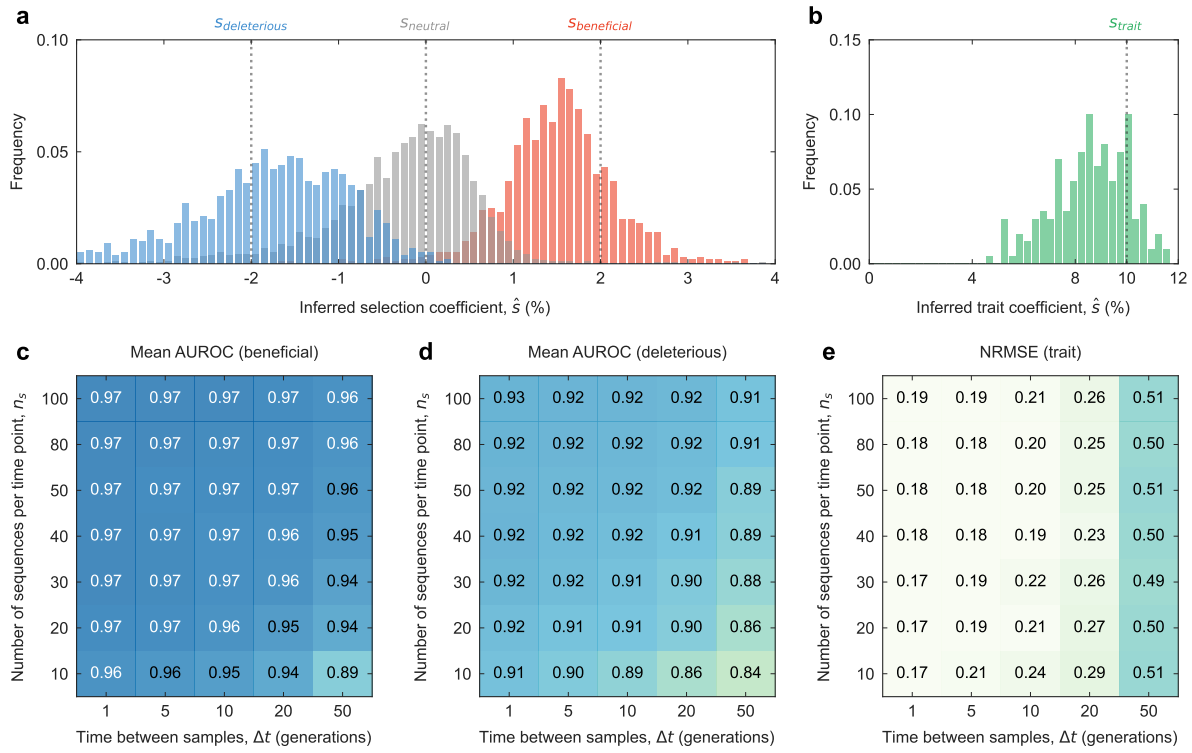
To answer this question, we inferred selection on mutant alleles and traits in simulations and in the HIV-1 data sets, with and without the inclusion of recombination. In general, we find that the effect of recombination on the inferred coefficients is small. This is reasonable, as recombination only affects the traits, which constitute a small portion of a sequence. For simulation data, there are three traits in a sequence length of 50; however, for experimental data, it is often a few epitopes contained within sequences that are thousands of base pairs long.

Simulation results (**Supplementary Fig. S7**) illustrate that the recombination term has negligible effects on the inference of individual selection coefficients. For trait terms, incorporating recombination does lead to small but noticeable improvements in inference. In HIV-1 data, the influence of the recombination term is small (**Supplementary Fig. S8**). Although the recombination rate in HIV-1 is relatively high, the scenarios under which recombination will lead to net change in trait frequencies are rare.

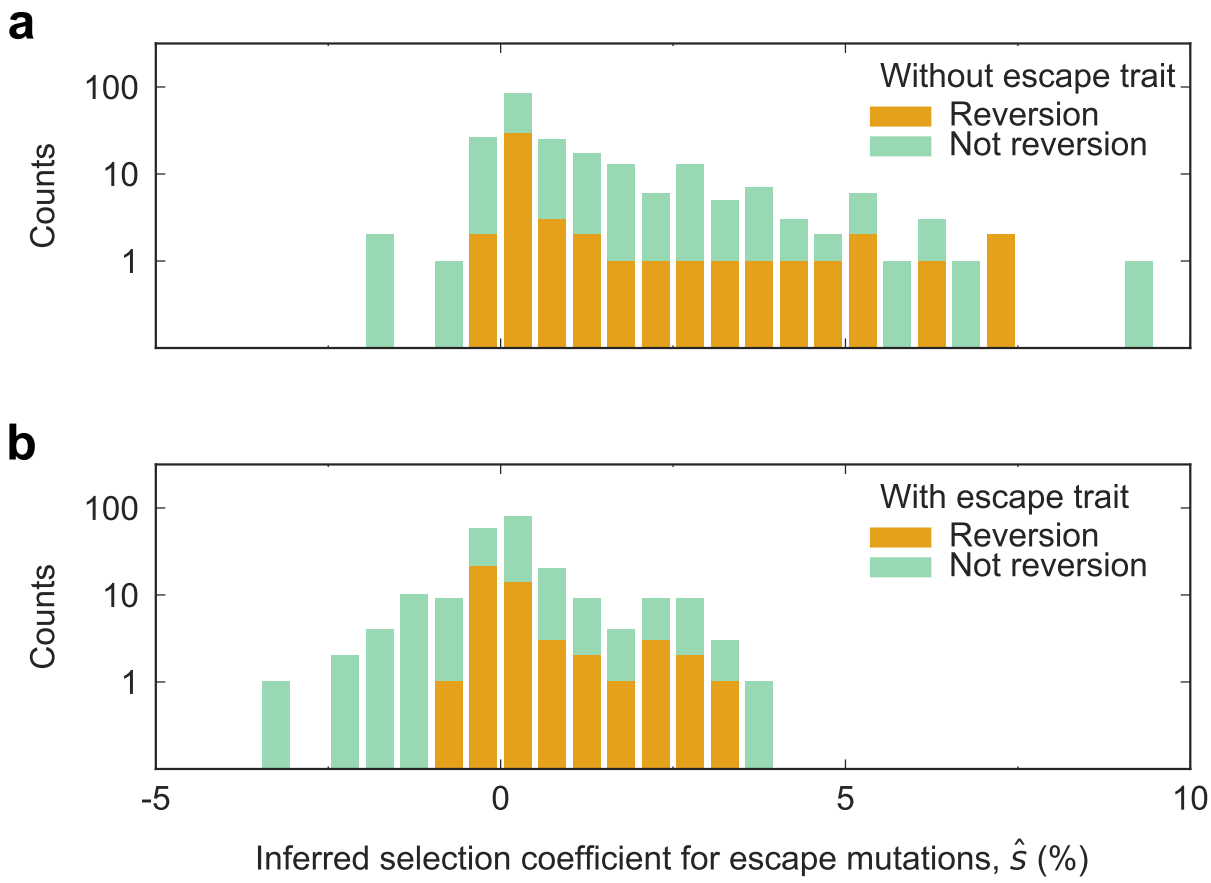
Data and code. Data and code used in our analysis is available in the GitHub repository: https://github.com/bartonlab/paper-binary-trait-inference. This repository also contains Jupyter notebooks that can be run to reproduce our figures and analysis.



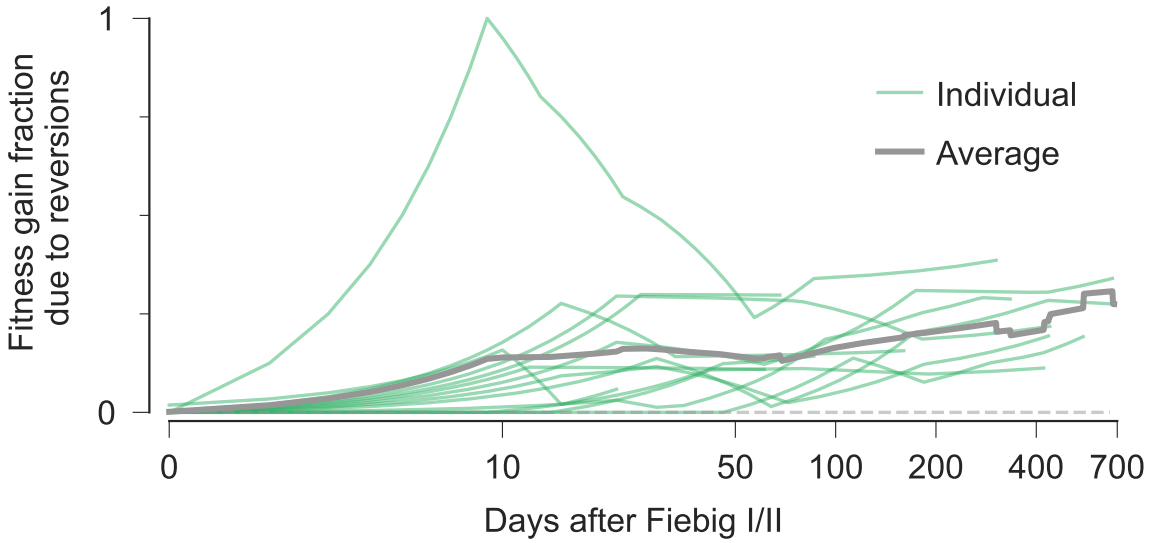
Supplementary Fig. S1. MPL recovers selection from data with time-varying selection coefficients for binary traits. a, Simulated mutant allele frequency trajectories. Outlined allele frequencies represent the alleles that affect one of the two traits. **b**, Trait frequencies and their contributing individual mutant allele frequencies (outlined in (**a**)) in the same simulation. The fitness contributions of individual mutations (**c**) and traits (**d**) that we infer are close to their true values. In **c**, error bars with black outlines represent alleles that affect the traits. The true values for the selection coefficient for individual loci are constant while those for binary traits vary over time. Inferred trait coefficients are close to the average values of time-varying coefficients over the course of the simulation. Distribution of (**e**) individual selection coefficient and (**f**) trait coefficient estimates across 100 replicate simulations. The true value for the selection coefficients for the binary traits shown here is the average value. Simulation parameters: $L = 50$ loci with two alleles at each locus (mutant and WT), ten beneficial mutants with $s = 0.02$, 30 neutral mutants with $s = 0$ and ten deleterious mutants with $s = -0.02$. We consider two binary traits, each with three contributing alleles and time-varying trait coefficients. The selection coefficients for these two binary traits go from 0 to 0.1 over the beginning 200 generations and then go down to 0 over the remaining 800 generations. Mutation probability per site per generation $\mu = 2 \times 10^{-4}$, recombination probability per site per generation $r = 2 \times 10^{-4}$, population size $N = 10^3$. The initial population contains all WT sequences, evolved over $T = 1000$ generations.



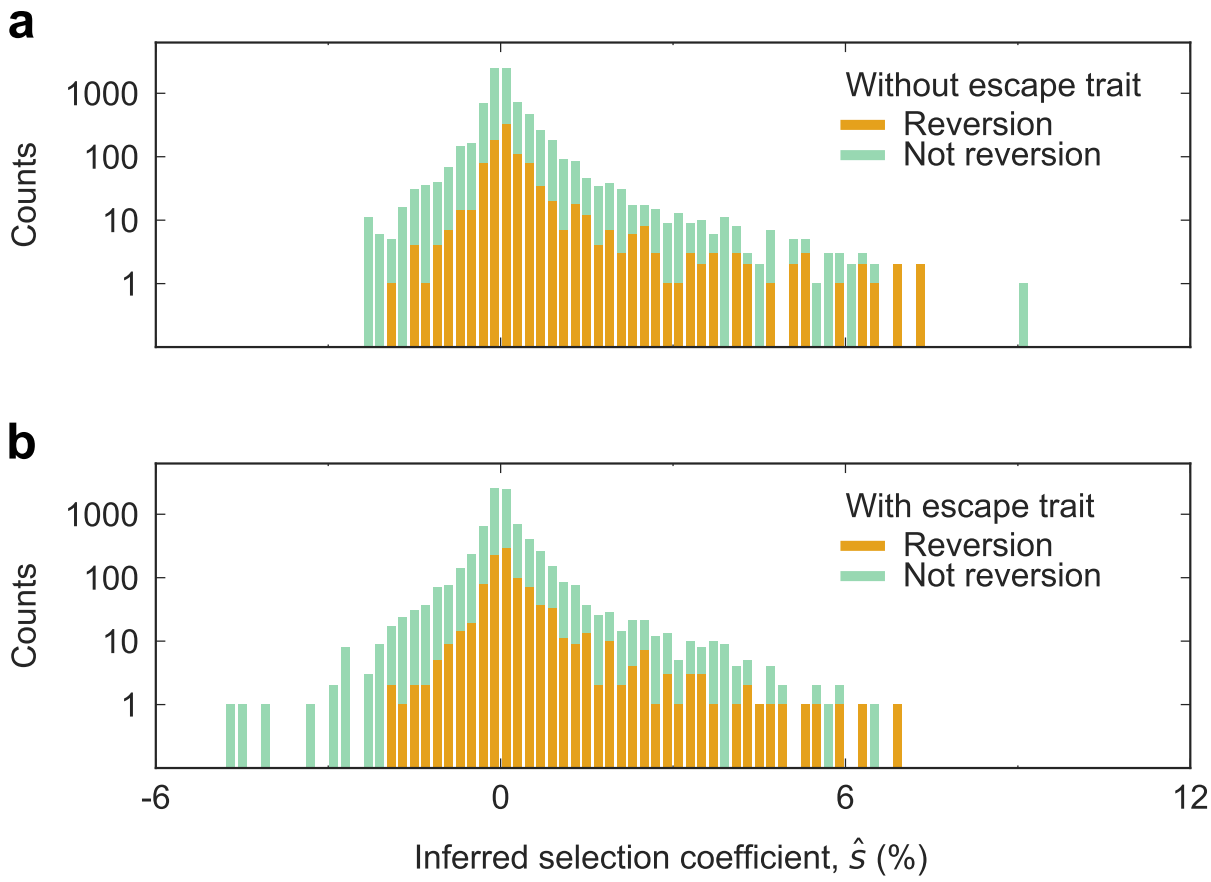
Supplementary Fig. S2. MPL recovers selection from complex dynamics even from limited data. Distribution of (**a**) individual selection coefficient and (**b**) trait coefficient estimates across 100 replicate simulations, using the same parameters as in **Fig. 1**. Our approach is robust to finite sampling constraints, as measured by the accurate classification of (**c**) beneficial and (**d**) deleterious mutants and (**e**) inference of trait coefficients, even when the number of sequences sampled per time point n_s is low and the spacing between time samples Δt is large. AUROC, area under the receiver operating characteristic; NRMSE, normalized root mean square error. Simulation parameters: $L = 50$ loci with two alleles at each locus (mutant and WT), ten beneficial mutants with $s = 0.02$, 30 neutral mutants with $s = 0$, and ten deleterious mutants with $s = -0.02$. We consider $\Lambda = 2$ trait groups, each with three contributing alleles and trait coefficients $s = 0.1$. Mutation probability per locus per generation $\mu = 2 \times 10^{-4}$, population size $N = 10^3$. The initial population is all wild type, evolved over $T = 1000$ generations.



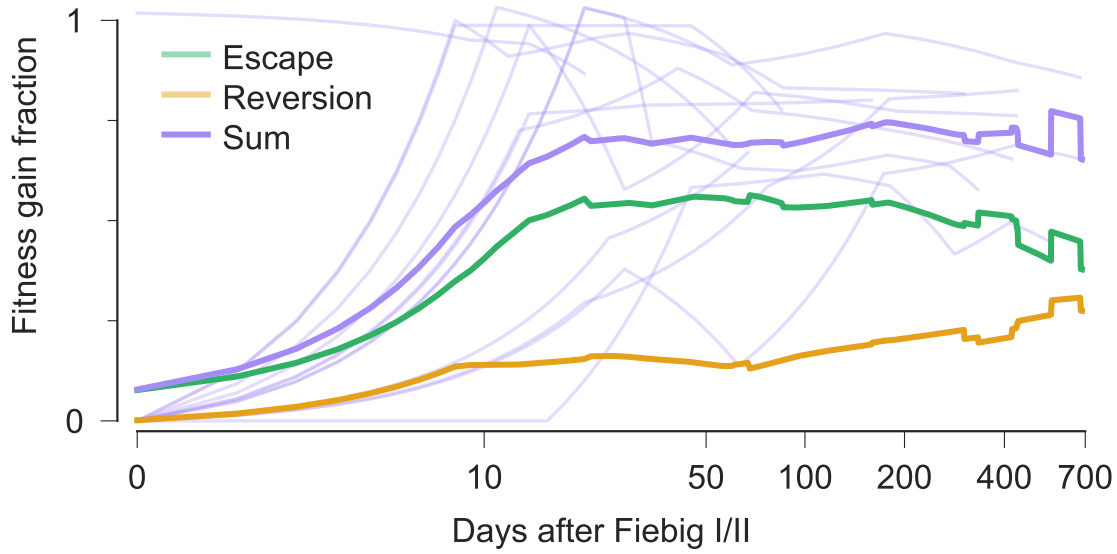
Supplementary Fig. S3. Inclusion of escape coefficients substantially shifts inferred selection coefficients for escape mutations toward more deleterious values. **a**, Distribution of inferred selection coefficients without escape traits. **b**, Distribution of inferred selection coefficients with escape traits. Reversions are slightly more likely to be inferred to be beneficial/less like to be deleterious than non-reversions.



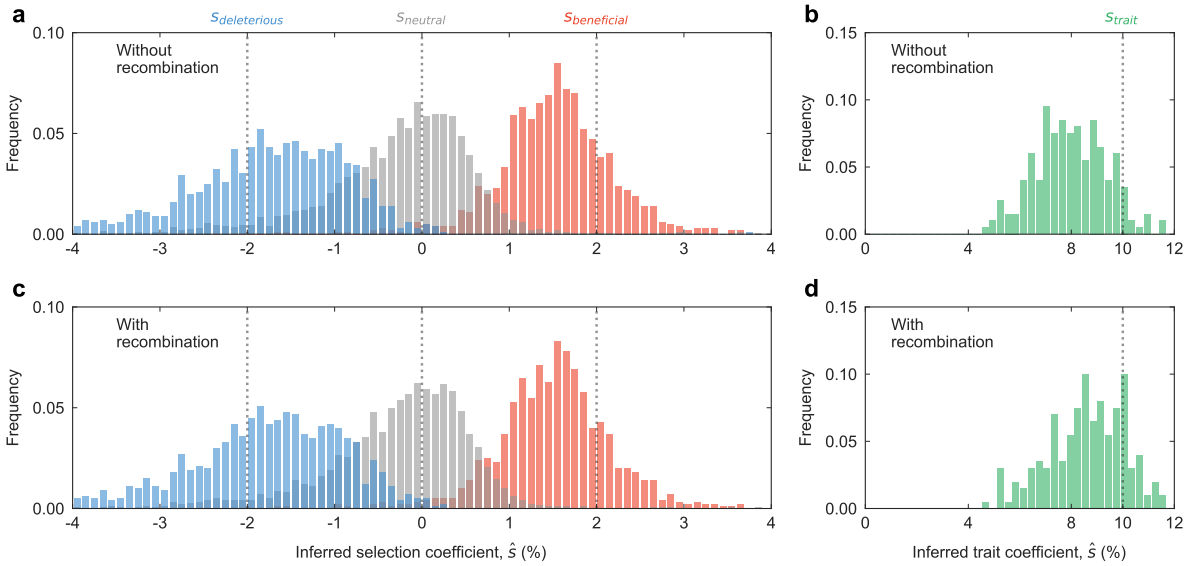
Supplementary Fig. S4. The contribution of reversions to intrahost HIV-1 fitness gains is significant and grows over time. Reversions refer to mutations that revert from the transmitted/founder (TF) variant to the nucleotide of the HIV-1 clade consensus sequence.



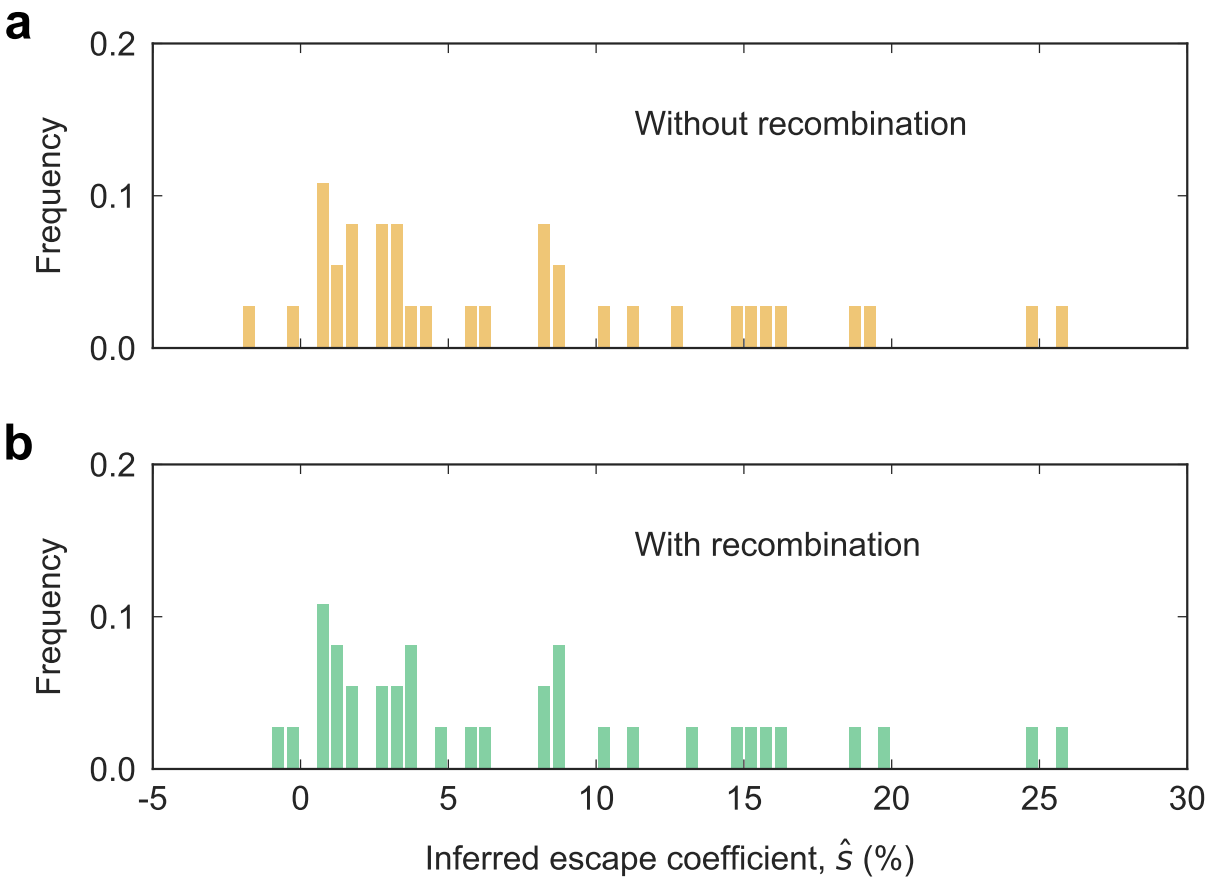
Supplementary Fig. S5. The distribution of inferred selection coefficients with and without the inclusion of escape traits. **a**, Distribution of inferred selection coefficients without escape traits. **b**, Distribution of inferred selection coefficients with escape traits. We note an enrichment in substantially beneficial mutations that are also reversions to the HIV-1 clade consensus sequence. While some large-effect mutations shift substantially with the inclusion of escape traits (i.e., escape mutations), the effect on the bulk of the inferred selection coefficients is minimal.



Supplementary Fig. S6. Intrahost HIV-1 fitness gains due to CD8$^+$ T cell escape mutations and reversions. In total, CD8$^+$ T cell escape (see **Fig. 2b**) and reversions (see **Supplementary Fig. S4**) make dominant contributions to HIV-1 fitness gains *in vivo*. Across the data sets from 13 people living with HIV-1 (PLHIV) that we studied, the fraction of fitness gains due to escape and reversions is around 75% on average. At earlier times, T cell escape plays a larger role, while the contribution of reversions grows steadily over time.



Supplementary Fig. S7. Effects of recombination on inferred selection coefficients and trait coefficients in 100 replicate simulations. The distribution of inferred (**a**) selection coefficients and (**b**) trait coefficients in simulations, using the same parameters as in **Fig. 1**, without including the effects of recombination in the estimator. **c**, **d**, Analogous distributions of inferred coefficients when recombination is included in the estimator. The effects of recombination on the inferred parameters are subtle, with the most prominent feature being a shift in the inferred trait coefficients toward the true value and away from zero.



Supplementary Fig. S8. Effects of recombination on inferred escape coefficients in HIV-1 data. Distribution of inferred escape coefficients (**a**) without and (**b**) with the inclusion of recombination in the estimator. Even though the effective recombination rate for HIV-1 is relatively high ($r \sim 1.4 \times 10^{-5}$, and even higher during acute infection when viral loads are high (10)), recombination events capable of altering trait frequencies are rare. The contribution of $\boldsymbol{R}_{\mathrm{fl}}$ to the inferred escape coefficients is thus small.

References

1. WJ Ewens, *Mathematical population genetics: theoretical introduction.* (Springer) Vol. 27, (2004).
2. V Mustonen, M Lässig, Fitness flux and ubiquity of adaptive evolution. *Proc. Natl. Acad. Sci.* **107**, 4248–4253 (2010).
3. MS Sohail, RH Louie, MR McKay, JP Barton, Mpl resolves genetic linkage in fitness inference from complex evolutionary histories. *Nat. biotechnology* **39**, 472–479 (2021).
4. B Lee, et al., Inferring effects of mutations on sars-cov-2 transmission from genomic surveillance data. *medRxiv* pp. 2021–12 (2022).
5. A Moumen, L Polomack, B Roques, H Buc, M Negroni, The hiv-1 repeated sequence r as a robust hot-spot for copy-choice recombination. *Nucleic acids research* **29**, 3814–3821 (2001).
6. TE Schlub, RP Smyth, AJ Grimm, J Mak, MP Davenport, Accurately measuring recombination between closely related hiv-1 genomes. *PLoS computational biology* **6**, e1000766 (2010).
7. RP Smyth, et al., Identifying recombination hot spots in the hiv-1 genome. *J. virology* **88**, 2891–2902 (2014).
8. H Song, et al., Tracking hiv-1 recombination to resolve its contribution to hiv-1 evolution in natural infection. *Nat. communications* **9**, 1928 (2018).
9. F Zanini, V Puller, J Brodin, J Albert, RA Neher, In vivo mutation rates and the landscape of fitness costs of hiv-1. *Virus evolution* **3**, vex003 (2017).
10. EV Romero, AF Feder, Elevated hiv viral load is associated with higher recombination rate in vivo. *Mol. Biol. Evol.* **41**, msad260 (2024).
11. MK Liu, et al., Vertical t cell immunodominance and epitope entropy determine hiv-1 escape. *The J. clinical investigation* **123** (2012).
12. RA Neher, T Leitner, Recombination rate and selection strength in hiv intra-patient evolution. *PLoS computational biology* **6**, e1000660 (2010).
13. JP Barton, et al., Relative rate and location of intra-host hiv evolution to evade cellular immunity are predictable. *Nat. communications* **7**, 11660 (2016).
14. MS Sohail, RH Louie, Z Hong, JP Barton, MR McKay, Inferring epistasis from genetic time-series data. *Mol. biology evolution* **39**, msac199 (2022).