Mean-field computational approach to HIV dynamics on a fitness landscape

Hanrong Chen[1¤], Mehran Kardar[2*]

1 Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, United States of America
2 Department of Physics, Massachusetts Institute of Technology, Cambridge, MA, United States of America

¤Current Address: Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA, United States of America
* kardar@mit.edu

Abstract

During acute infection by the human immunodeficiency virus (HIV), the intra-host viral population accumulates mutations due to selection by host T cell responses. The timescales at which HIV mutations emerge vary widely, correlating with their diversity in the global population of hosts, and with the relative strengths at which different regions of the HIV sequence are targeted by the host. In recent years, "fitness landscapes" of HIV proteins have been estimated from this global diversity, capturing residue-specific fitness costs and epistatic interactions between residues, and have been shown to predict the locations and relative timescales of escape mutations arising in hosts with known T cell responses. However, existing computational approaches to HIV dynamics that make use of realistic fitness landscapes are limited to fixed-population-size stochastic simulations, and extending other dynamical approaches to capture full sequence-level detail is computationally nontrivial. In this paper, we introduce and study an alternative approach for simulating HIV dynamics given a fitness landscape, which we designate the evolutionary mean-field (EMF) method. EMF is the high-recombination-rate limit of a model of HIV replication and mutation, which we justify using methods from statistical physics. EMF outputs a set of time-dependent "effective fitnesses" that crucially depend on epistatic interactions and the sequence background, as well as predictions of the frequencies of mutation at each HIV residue over time. As a proof of principle, we apply this method to the dynamics of the p24 gag protein infecting a host whose T cell responses are known. Specifically, we show how fitness costs and epistatic interactions in the fitness landscape, the relative strengths of T cell responses, and the HIV sequence background, impact the locations and time course of HIV escape mutations, consistent with previous work. We also describe features of longer-term dynamics, specifically reversions, in terms of the effective fitnesses yielded by EMF, and quantify the mean fitness and entropy of the intra-host population over time. Finally, we develop a stochastic population dynamics version of EMF, extending prior stochastic approaches to a time-varying population size that crucially depends on the fitness of strains existing in the intra-host population at each time. The EMF approach offers a framework for understanding features of HIV dynamics in terms of effective fitnesses, and allows for a more detailed study of how the fitness landscape and sequence background impact both the evolutionary and population dynamics of HIV, in a computationally tractable way.

Author summary

As fitness landscapes of HIV proteins become more accurately known, it may become possible to more faithfully predict the locations and timescales of HIV mutations arising in a host with known immune responses, which may inform the design of vaccine immunogens. However, existing approaches for computing HIV dynamics given a fitness landscape are limited to fixed-population-size stochastic simulations. Here, we present an alternative approach that we designate the evolutionary mean-field (EMF) method. It takes as input an HIV fitness landscape and the locations and strengths of host immune responses, and outputs a set of time-dependent "effective fitnesses" and frequencies of mutation at each HIV residue over time. EMF is the high-recombination-rate limit of a model of HIV replication and mutation, which we derive using methods from statistical physics. We apply EMF on an example to show how fitness costs and epistatic interactions in the fitness landscape, the relative strengths of host immune responses, and the HIV sequence background, impact the locations and time course of HIV mutations. We also develop a stochastic population dynamics version of EMF where population size changes crucially depend on the fitness of strains existing in the population at each time. EMF makes quick predictions of the dynamics of HIV mutations through the effective fitnesses, crucially taking into account the fitness landscape and sequence background, and enables more detailed studies of how these affect the evolutionary and population dynamics of HIV, in a computationally tractable way.

Introduction

During acute infection by the human immunodeficiency virus (HIV), the virus replicates rapidly and the plasma viral load (concentration of HIV RNA particles in the blood plasma), a measure of magnitude of HIV infection, rises exponentially to a peak 2–3 weeks post-infection [1]. Host cytotoxic T lymphocyte (CTL) responses are detectable just before this peak and expand as the viral load declines [1–3]. CTLs kill HIV-infected host cells through the recognition of HIV-derived peptides 8–10 amino acids long (called epitopes) that are presented on the surface of infected cells bound to major histocompatibility complex (MHC) class I molecules [4]. While CTLs partially control HIV in an untreated host [5], they are unable to clear the infection. Mutations occur during HIV replication due to error-prone reverse transcription [6], and those lying within epitopes allow such mutants to escape epitope-specific CTL recognition. Thus, CTL-driven selection acts on the mutating intra-host HIV population and selects for the emergence of escape mutants over time, which can be measured by sequencing HIV in blood drawn from infected hosts [3, 7–9].

The timescales at which HIV escape mutations emerge in the intra-host population vary widely, ranging from days to years [1–3]. Why is this so, and how does this depend on properties of the virus and/or host immune response? Liu et al. [3] found that the rate of virus escape at an epitope correlates with the entropy of epitope sequences in the global population of hosts infected by the same group or subtype of HIV; they also found that the relative strength of CTL responses targeting each epitope (called the vertical immunodominance) plays an important role [3]. These results suggest that different HIV strains have different levels of replicative *fitness*: residues where mutations incur large fitness costs tend to be more conserved in the global population of hosts and also tend to acquire escape mutations more slowly; also, CTL responses impose additional fitness costs that are abrogated by escape mutations, and higher vertical immunodominance corresponds to larger costs imposed at an epitope. Using global entropy (or diversity) as an indicator for fitness is also supported by mutations

toward the group or subtype consensus residue (called reversions) that tend to occur when HIV is transmitted to an MHC-mismatched host (who presents different epitopes from the donor) [10, 11] and continue over years of intra-patient evolution [9]. In recent years, residue-specific fitness costs of HIV mutants have been measured in cell culture [12–15] and from genetic variation within individual hosts [16], finding a correlation with site entropy among global HIV strains of the same group or subtype (although less conserved residues also tend to be more commonly targeted by CTLs [16], and low-fitness-cost mutants may sometimes be rarely observed globally [14, 15]). Apart from residue-specific fitness costs, epistatic interactions between residues were also shown to be important in particular for the HIV gag [17–21] and nef [22] proteins, and this should be reflected in pairwise correlations in the global diversity of sequences. In recent years, fitness "landscapes" of various HIV proteins were inferred from this global diversity [12, 13, 23], providing effective descriptions of intrinsic HIV fitness as a function of amino acid sequence that include epistatic interactions between residues. Using such inferred landscapes, Barton et al. [24] performed *in silico* simulations of HIV infecting patients with known CTL responses, and found a good correlation in the locations and timescales of escape mutations arising *in silico* and measured in the patients [24]. More broadly, as characterizations of HIV fitness landscapes become more accurate, including epistatic interactions, this may enable more accurate predictions of the dynamics of HIV mutations arising in a given host.

In this paper, we focus on methods for simulating intra-host HIV dynamics given a residue-specific fitness landscape such as those inferred previously [12, 13, 23]. Doing so is nontrivial as the presence of epistatic interactions deem it insufficient to simulate just one site or one epitope; the sequence background matters (as we will show later). Currently available approaches for simulating the dynamics of full HIV protein sequences given a fitness landscape are limited to fixed-population-size stochastic simulations [24]. A fixed population size does not reflect its changes by orders of magnitude during acute infection [1], which may affect the dynamics of mutations that arise. Also, a stochastic method requires performing many (10^3 in [24]) simulations in order to start making predictions about the locations and rates of HIV escape [24], and after the fact it may not be clear why certain mutations tend to occur and not others. Compartmental models of HIV dynamics that consider multiple HIV strains and multiple CTL responses [25–28] could be extended to have one set of equations for each HIV strain in the population, but this would require a very large number of equations as the intra-host population expands and diversifies.

Here, we develop an alternative approach for the quick but approximate computation of HIV dynamics given an HIV fitness landscape and host CTL responses, that we designate the evolutionary mean-field (EMF) method. The "mean-field" in EMF does not refer to chemical species reacting according to their mean concentrations (as in compartmental models), but to a class of approximations in statistical physics. Specifically, we make use of a mapping of viral replication and mutation dynamics to a statistical physics model (first introduced in Refs [29, 30]), a mean-field approximation of which gives a set of "effective fields" or fitnesses describing the tendency of each HIV residue to mutate, crucially taking into account epistatic interactions and the sequence background, leading to estimates of the frequencies of mutation at each residue over time. Similar mean-field equations were obtained in Neher and Shraiman [31] in the limit of high recombination rate, which is precisely the biological interpretation of EMF. While EMF is manifestly an approximation, in this limit the dynamics remain computationally tractable, and it nevertheless enables us to study new questions regarding how genetic-level attributes of HIV and the host immune response influence the dynamics. As a proof of principle, we apply EMF on the dynamics of the p24 gag protein infecting a host whose CTL responses are known, and we show how fitness costs

and epistatic interactions within the fitness landscape, vertical immunodominance of CTL responses, and the HIV sequence background, impact the locations and relative timescales of HIV escape mutations. In particular, some of these predictions rely crucially on the epistatic interactions, which are automatically encoded in the effective fitnesses yielded by EMF. We also describe reversions and potential compensatory mutations in terms of these effective fitnesses, and quantify the mean fitness and entropy of the intra-host population over time. To validate these results, in S2 Appendix we perform fixed-population-size stochastic simulations (as in Ref [24]) for this example.

In the final section of this paper, we present a stochastic population dynamics version of EMF, where population size changes depend crucially on the fitness of strains in the intra-host population at each time, extending prior stochastic approaches which have a fixed population size [24, 32] or a time-varying one that is externally defined [33]. We show that the locations of escape mutations are now stochastic, but the deterministic EMF method provides good predictions on average. This method also produces a characteristic exponential rise and fall of the HIV population size that is qualitatively consistent with the plasma viral load during acute infection. We end by discussing limitations, uses, and further extensions of EMF, including how to incorporate explicit target cell and CTL clone populations, like other compartmental models of HIV dynamics [25–28]. The EMF approach offers a framework for describing various features of HIV dynamics in terms of effective fitnesses, enables more detailed study of how the fitness landscape and sequence background affect the evolutionary and population dynamics of HIV in a computationally tractable way, and provides quick dynamical predictions that can be checked against other, more computationally intensive methods implementing a finite recombination rate.

Methods

The fitness landscape of an HIV protein

Following prior work [12, 13, 24], we model the fitness landscape of an HIV protein as follows: We describe HIV strains by their amino acid sequence. Here we assume a binary approximation (done in [12] but not [13, 24]) and describe an HIV sequence as $S = (s_1, \cdots, s_L)$, where L is the length of the protein, and each $s_i = 0, 1$ where $s_i = 0$ represents the consensus amino acid at position i and $s_i = 1$ represents any other residue. This approximation is reasonable for a protein which is highly conserved (such as p24 gag [12]); taking s_i to be one of four nucleotides or one of twenty amino acids is a straightforward extension, but is beyond the scope of this paper. We consider the intrinsic fitness of HIV strain S to obey the following form:

$$F_{\text{in}}(S) = F_0 + \sum_{i=1}^{L} h_i s_i + \sum_{i<j} J_{ij} s_i s_j, \tag{1}$$

where F_0 is the intrinsic fitness of the consensus strain, $\{h_i\}$ capture fitness costs of mutating from the consensus residue, and $\{J_{ij}\}$ capture pairwise epistatic interactions between residues[1]. In Ref [24], the values of these parameters are related to the prevalence landscape of the protein (see Parameter estimation).

During intra-host infection, the fitness of HIV strain S is modified from its intrinsic value by host CTL pressure. Following Ref [24], we describe the effect of CTL-mediated killing of HIV-infected cells by an effective fitness cost b_ε, assigned to HIV strains S with no mutation in epitope ε. This fitness cost is abrogated if S has at least one

[1] Our methods can easily be extended to fitness landscapes with higher-order interactions; indeed, CTL epitopes are an example of a higher-order interaction (see below).

mutation within ε, which tends to nullify CTL recognition. For a set of epitopes $\{\varepsilon\}$ at sites $(i_\varepsilon, \cdots, i_\varepsilon + l_\varepsilon - 1)$ (i.e., beginning at site i_ε of S and of length l_ε amino acids), and targeted epitope sequences $(0, \cdots, 0)$, the within-host fitness of HIV strain S becomes

$$F_{\text{host}}(S) = F_{\text{in}}(S) - \sum_{\varepsilon} b_\varepsilon (1 - s_{i_\varepsilon}) \cdots (1 - s_{i_\varepsilon + l_\varepsilon - 1}). \tag{2}$$

(Replace $1 - s_{j_\varepsilon} \to s_{j_\varepsilon}$ if epitope ε has $s_{j_\varepsilon} = 1$.) In general, b_ε is time-dependent so $F_{\text{host}}(S)$ is time-dependent.

Model of HIV dynamics

Within a host, an HIV particle replicates by infecting a host CD4$^+$ cell and reverse transcribing its RNA genome into DNA. HIV DNA is then integrated into host DNA and new copies of the mutated virus are produced by the target cell. The ability of an HIV strain to replicate and infect new target cells is effectively described by its fitness F_{host} (see above). Here, we model the process of replication using a discretized-time Markov model (see Fig 1). In one replication cycle (which takes roughly 1 day [34] [2]), the effective number of offspring[3] of HIV strain S is $e^{F(S)}$ where $F = F_{\text{host}}$. Reverse transcription is an error-prone process and introduces random point mutations within the HIV genome with a probability that depends on nucleotide-to-nucleotide substitution rates [6,16]. In the following, we consider a mutation rate μ that is site and state-independent; we show in S1 Appendix that our results are similar when considering mutation rates that are nucleotide-dependent.

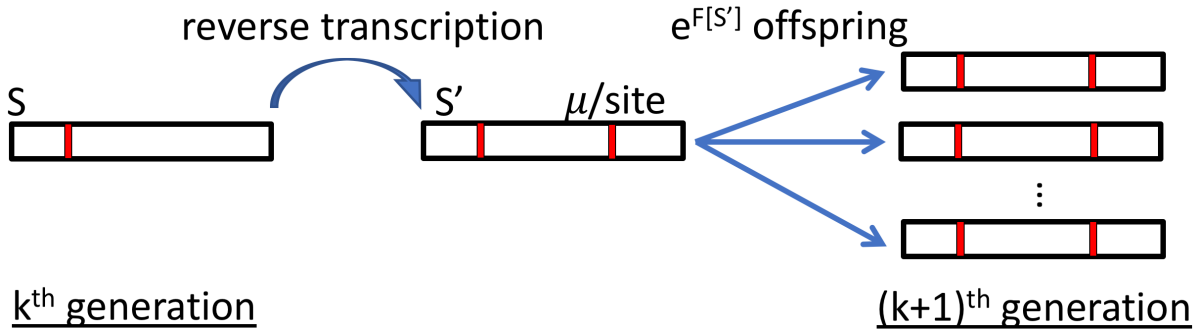

Fig 1. Model of HIV replication and mutation. In one replication cycle, HIV sequence S reverse transcribes its genome, introducing random point mutations and producing sequence S'. Sequence S' then produces a number of offspring given by its fitness, $e^{F(S')}$. See Eq (5) for the transition matrix describing this process.

First, let us write down the equations for replication and mutation of a single site s_i. Suppose $s_i = 0$ has fitness F_0 and $s_i = 1$ has fitness $F_0 + h_i$. Also suppose that $s_i = 0$ mutates to $s_i = 1$ or vice-versa with probability μ/generation. Let the frequency of $s_i = 1$ in the population of size $N_i^{(k)}$ at time k be $m_i^{(k)}$. The population sizes of $s_i = 0$

[2] Discretizing time more finely requires rescaling the replication and mutation rates appropriately (see Discussion).

[3] In this paper, we do not explicitly model infection of target cells or target cell dynamics. In the Discussion, we discuss extensions of the EMF method that include explicit consideration of a target cell population.

and $s_i = 1$ at time $(k+1)$ are thus

$$\begin{aligned}
&\begin{pmatrix} e^{F_0} & 0 \\ 0 & e^{F_0+h_i} \end{pmatrix} \begin{pmatrix} 1-\mu & \mu \\ \mu & 1-\mu \end{pmatrix} \begin{pmatrix} 1-m_i^{(k)} \\ m_i^{(k)} \end{pmatrix} N_i^{(k)} \\
=&e^{F_0} \begin{pmatrix} (1-\mu)(1-m_i^{(k)}) + \mu m_i^{(k)} \\ e^{h_i}[\mu(1-m_i^{(k)}) + (1-\mu)m_i^{(k)}] \end{pmatrix} N_i^{(k)},
\end{aligned} \tag{3}$$

giving a total (increased or decreased) population size at time $(k+1)$ of $N_i^{(k+1)}$, with ratio

$$\tilde{Z}_i^{(k+1)}(m_i^{(k)}) \equiv \frac{N_i^{(k+1)}}{N_i^{(k)}} = e^{F_0}\{(1-\mu)(1-m_i^{(k)}) + \mu m_i^{(k)} + e^{h_i}[\mu(1-m_i^{(k)}) + (1-\mu)m_i^{(k)}]\}. \tag{4}$$

Note that here mutations occur before replication (Fig 1) because reverse transcription occurs before forming new virus particles (and this order is important for deriving the EMF equations below), but a model with replication before mutation can be thought of as having one fewer round of mutations at the beginning, and hence would produce essentially the same results after many time steps (not shown).

Thus, for sequences of length L replicating according to a fitness landscape $F = F_{\text{host}}$, the number of offspring with sequence $S^{(k+1)}$ produced by a single sequence $S^{(k)}$ in one generation is given by

$$\langle S^{(k+1)}|T|S^{(k)}\rangle = e^{F(S^{(k+1)})} \prod_{i=1}^{L} (1-\mu)^{\delta_{s_i^{(k)},s_i^{(k+1)}}} \mu^{1-\delta_{s_i^{(k)},s_i^{(k+1)}}}, \tag{5}$$

where $\delta_{s_i^{(k)},s_i^{(k+1)}}$ is the Kronecker delta function. The $2^L \times 2^L$ transition matrix T is not a Markov transition probability matrix (which would require $\sum_{S'}\langle S'|T|S\rangle = 1$), but of a model of the time-varying population size. The state at each time is a vector $\mathbb{P}$ of populations of the 2^L sequences S. Suppose we know the within-host fitness landscape F_{host} and the initial intra-host population $\mathbb{P}(0)$. Our goal is to predict $\mathbb{P}$ at future times, specifically to answer the following two kinds of questions:

1. What is the steady state at long times? Which escape mutations arise? Are there reversions or compensatory mutations?

2. What are the dynamics towards the steady state? What is the order and timescale of mutations? Are there transient dynamics?

The most direct method of solution is to find $\mathbb{P}(n) = T^n\mathbb{P}(0)$, from which we can compute, for example, the statistics of mutations at site i over time. This involves, at time k, finding the marginal frequencies

$$m_i(k) = \frac{\sum_{s_1}\cdots\sum_{s_L}\mathbb{P}(k)s_i}{\sum_{s_1}\cdots\sum_{s_L}\mathbb{P}(k)}. \tag{6}$$

Equipped with this information, we can for example predict the relative probabilities of each escape mutation within a targeted epitope.

This direct method is computationally intractable. For an HIV protein of length $L = 200$, there are $2^{200} \approx 10^{60}$ possible sequences, so diagonalizing or matrix decompositions of T aren't feasible in general. In practice, if the population size is not prohibitively large or if $\mathbb{P}$ is sparse, then all sequences existing in the population may be tracked individually. In particular, if $\mu N \ll 1$ (where $N = ||\mathbb{P}||$ is the population size),

successive selective sweeps occur and the population may be described by a dominant strain at each time [35]. However, this is not the regime HIV is in during the early stages of acute infection, when the intra-host population accumulates diversity through mutations that are selected upon by strong CTL pressure (or later in infection, when the population is yet more diverse). A commonly used method when $\mu N \gtrsim 1$ is to perform stochastic simulations with fixed population size N. In S2 Appendix, we perform such (Wright–Fisher) simulations to compute the dynamics for an example considered in Sec Results, following those by Barton et al. [24].

The evolutionary mean-field (EMF) method

Here, we develop an alternative approach to find the locations and dynamics of HIV mutations given a fitness landscape and host CTL responses. We first exploit a mapping of the dynamical model presented above to a statistical physics model (first introduced by Leuthausser [29, 30]). Using the identity $\delta_{s_i,s_i'} = \frac{1}{2}[1 + (1-2s_i)(1-2s_i')]$, we write Eq (5) in the form $e^{H(S^{(k+1)},S^{(k)})}$ with energy (Hamiltonian) given by

$$H(S^{(k+1)}, S^{(k)}) = F(S^{(k+1)}) + K \sum_{i=1}^{L} (1-2s_i^{(k)})(1-2s_i^{(k+1)}), \tag{7}$$

where $K \equiv \frac{1}{2}\log\frac{1-\mu}{\mu}$ (and we have omitted a constant term $\frac{L}{2}\log(1-\mu)\mu$ for notational clarity). The number of offspring produced by one HIV particle with sequence $S^{(k)}$ in one generation is thus given by

$$Z^{(k+1)}(S^{(k)}) \equiv \sum_{\{S^{(k+1)}\}} \langle S^{(k+1)}|T|S^{(k)}\rangle = \sum_{\{S^{(k+1)}\}} e^{H(S^{(k+1)},S^{(k)})}. \tag{8}$$

While $H(S^{(k+1)}, S^{(k)})$ depends on both $S^{(k)}$ and $S^{(k+1)}$, the 'partition function' $Z^{(k+1)}$ sums over the 2^L possible sequences $\{S^{(k+1)}\}$. Extending to multiple generations, the resulting recursive structure is shown in Fig 2, left, with sites (spins) within each generation interacting according to $F(S)$, and spins along the time dimension having nearest-neighbor couplings K [29, 30]. The partition function of this model precisely gives the total number of offspring produced by a sequence $S^{(0)}$ over n generations.

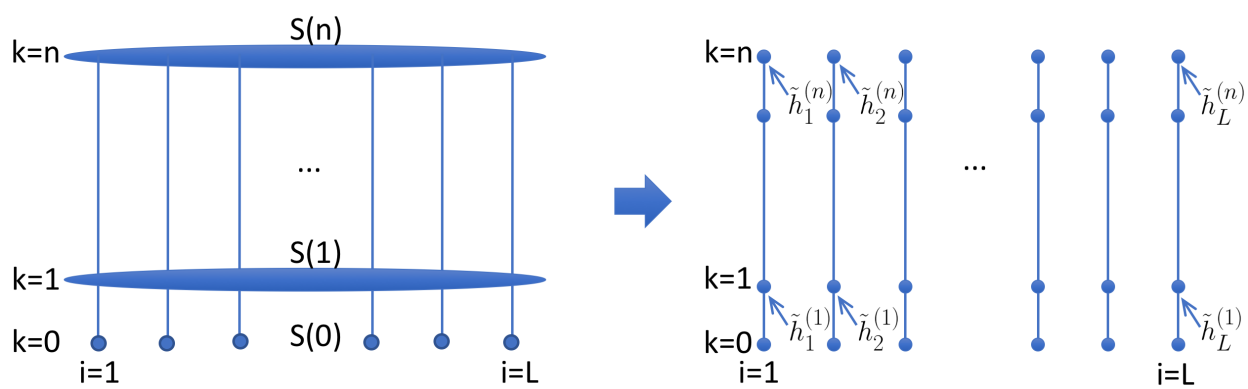

Fig 2. The evolutionary mean-field approximation. Left: A statistical physics model with Hamiltonian $H(S^{(n)}, \cdots, S^{(0)})$ (generalizing Eq (7) to multiple generations). Right: A schematic of the evolutionary mean-field (EMF) approximation. The "effective fitnesses" $\tilde{h}_i^{(k)}$ are time-dependent, even for a fitness landscape F that is constant in time.

To set up the evolutionary mean-field (EMF) approximation, consider a model of

noninteracting spins, with Hamiltonian

$$\tilde{H}(S^{(k+1)}, S^{(k)}) = \sum_{i=1}^{L} \Big[\tilde{h}_i^{(k+1)} s_i^{(k+1)} + K(1-2s_i^{(k)})(1-2s_i^{(k+1)})\Big], \tag{9}$$

where the "effective fields" or fitnesses $\tilde{h}_i^{(k+1)}$ at sites $i = 1, \cdots, L$ are to be determined. While the parameters of F_{host} do not depend on time apart from $b_\varepsilon(k)$, here the $\tilde{h}_i^{(k+1)}$ depend on the fitness landscape in a nontrivial way, as we will show below. In one iteration of the approximate model, we obtain

$$\tilde{Z}^{(k+1)}(S^{(k)}) = \prod_{i=1}^{L} \big[e^{\tilde{h}_i^{(k+1)} \mp K} + e^{\pm K}\big], \tag{10}$$

where the upper sign is for $s_i^{(k)} = 0$ and the lower sign is for $s_i^{(k)} = 1$ (cf. Eq (4) with $m_i^{(k)} = 0$ or 1, noting that $e^K = \sqrt{\frac{1-\mu}{\mu}}$). We want to find $\{\tilde{h}_i^{(k+1)}\}$ such that $Z(S^{(k)})$ and $\tilde{Z}(S^{(k)})$ are as "close" as possible. We make use of Gibbs' inequality [36], which states that the Kullback–Leibler divergence of two probability distributions P and $\tilde{P}$ is nonnegative, i.e., $D_{KL}(\tilde{P}||P) \geq 0$. Using the forms $P(\cdot) = e^{H(\cdot)}/Z$ and $\tilde{P}(\cdot) = e^{\tilde{H}(\cdot)}/\tilde{Z}$, Gibbs' inequality becomes

$$\log Z^{(k+1)} \geq \log \tilde{Z}^{(k+1)} + \langle H(S^{(k+1)}, S^{(k)}) - \tilde{H}(S^{(k+1)}, S^{(k)})\rangle_\sim, \tag{11}$$

where $\langle f\rangle_\sim \equiv \sum_x f(x)\tilde{P}(x)$ is the expectation value with respect to the approximate model. (We note here that Shekhar et al. [37] used Eq (11) in similar work to study the relation between the prevalence and intrinsic fitness landscapes of HIV; they used a different $\tilde{H}$ involving F_{in} instead of noninteracting spins, and their resulting equations do not reduce the computational complexity of the solution (which was not their goal) because their $\tilde{H}$ is still an interacting model.)

For the within-host fitness landscape, Eq (2),

$$\langle H - \tilde{H}\rangle_\sim = \sum_{i=1}^{L} \big(h_i - \tilde{h}_i^{(k+1)}\big) m_i^{(k+1)} + \sum_{i<j} J_{ij} m_i^{(k+1)} m_j^{(k+1)} - \sum_{\varepsilon} b_\varepsilon \prod_{j\in\varepsilon} (1 - m_j^{(k+1)}), \tag{12}$$

where $\langle s_i^{(k+1)} s_j^{(k+1)}\rangle = m_i^{(k+1)} m_j^{(k+1)}$ because sites are decoupled in the approximate model, and $m_i^{(k+1)} \equiv \langle s_i^{(k+1)}\rangle_\sim$. Hence, extremizing the RHS of Eq (11) w.r.t. $\tilde{h}_i^{(k+1)}$ leads to

$$\begin{aligned} 0 &= \frac{\partial}{\partial \tilde{h}_i^{(k+1)}} \Big[\log \tilde{Z}^{(k+1)} + \langle H - \tilde{H}\rangle_\sim\Big] \\ &= \cancel{m_i^{(k+1)}} - \cancel{m_i^{(k+1)}} + \Big[\frac{\partial}{\partial \tilde{h}_i^{(k+1)}} m_i^{(k+1)}\Big]\Big[h_i - \tilde{h}_i^{(k+1)} + \sum_{j\neq i} J_{ij} m_j^{(k+1)}\Big], \end{aligned} \tag{13}$$

for sites i outside of epitopes giving

$$\tilde{h}_i^{(k+1)} = h_i + \sum_{j\neq i} J_{ij} m_j^{(k+1)}, \tag{14}$$

and

$$\begin{aligned} 0 = \cancel{m_i^{(k+1)}} - \cancel{m_i^{(k+1)}} &+ \Big[\frac{\partial}{\partial \tilde{h}_i^{(k+1)}} m_i^{(k+1)}\Big] \\ &\times \Big[h_i - \tilde{h}_i^{(k+1)} + \sum_{j\neq i} J_{ij} m_j^{(k+1)} - b_\varepsilon \prod_{j\in\varepsilon\setminus i} (1 - m_j^{(k+1)})\Big], \end{aligned} \tag{15}$$

for sites i within epitope ε giving

$$\tilde{h}_i^{(k+1)} = h_i + \sum_{j \neq i} J_{ij} m_j^{(k+1)} - b_\varepsilon \prod_{j \in \varepsilon \setminus i} (1 - m_j^{(k+1)}). \tag{16}$$

(We have again assumed that the targeted epitope sequence is $(0, \cdots, 0)$; replace $1 - m_j^{(k+1)} \to m_j^{(k+1)}$ if it contains a site j with $s_j = 1$.) The $m_i^{(k+1)} \equiv \langle s_i^{(k+1)} \rangle_\sim$ are the frequencies of mutations within the approximate model. For $s_i^{(k)} = 0$ or 1, the frequency of mutation at time $k+1$ is

$$m_i^{(k+1)} = \frac{e^{\tilde{h}_i^{(k+1)} \mp K}}{e^{\pm K} + e^{\tilde{h}_i^{(k+1)} \mp K}}, \tag{17}$$

where again, the upper sign is for $s_i^{(k)} = 0$ and the lower sign is for $s_i^{(k)} = 1$. For generic $0 \leq m_i^{(k)} \leq 1$, we obtain

$$m_i^{(k+1)} = \frac{e^{\tilde{h}_i^{(k+1)}} [\mu(1 - m_i^{(k)}) + (1-\mu) m_i^{(k)}]}{\tilde{Z}_i^{(k+1)} \left(m_i^{(k)}; \tilde{h}_i^{(k+1)} \right)}, \tag{18}$$

(cf. Eqs (3) and (4)).

Equations (14), (16), and (18) form a set of self-consistent equations that relate the effective fitnesses $\tilde{h}_i^{(k+1)}$, $i = 1, \cdots, L$, at each time $k+1$ to the parameters of the fitness landscape, the CTL response $b_\varepsilon(k+1)$, and the frequencies of mutations $m_j^{(k+1)}$ at other sites $j \neq i$ at the same time $k+1$. While computing the marginal frequencies at each time by Eq (6) involves a sum over 2^L terms, here Eq (18) contains a sum of no more than two terms in the denominator, however requiring simultaneous solution with Eqs (14) and (16) (because $\tilde{h}_i^{(k+1)}$ depends on $m_j^{(k+1)}$ for $j \neq i$ which in turn depend on $\tilde{h}_j^{(k+1)}$). A method of solution is to iterate between $\{\tilde{h}_i^{(k+1)}\}$ and $\{m_i^{(k+1)}\}$ until convergence (not shown). However, if the $\{m_i^{(k)}\}$ do not change drastically with time, a computationally more efficient implementation is to replace $m_i^{(k+1)}$ with $m_i^{(k)}$ in Eqs (14) and (16). Thus, given a set of initial mutant frequencies $\{m_i^{(0)}\}$, we "shoot" forward in time k, iterating between $\{\tilde{h}_i^{(k)}\}$ and $\{m_i^{(k)}\}$ for larger and larger k:

$$\{m_i^{(0)}\} \xrightarrow{\text{Eqs (14) and (16)}} \{\tilde{h}_i^{(1)}\} \xrightarrow{\text{Eq (18)}} \{m_i^{(1)}\} \xrightarrow{\text{Eqs (14) and (16)}} \{\tilde{h}_i^{(2)}\} \to \cdots. \tag{19}$$

Finally, since $Z^{(k+1)}$ (Eq (8)) gives the ratio of population sizes at times $k+1$ and k, we may approximate the dynamics of population sizes using the RHS of Eq (11):

$$\log Z^{(k+1)} \geq \sum_{i=1}^{L} \log \tilde{Z}_i^{(k+1)} + F_{\text{host}}(m_1^{(k+1)}, \cdots, m_L^{(k+1)}) - \sum_{i=1}^{L} \tilde{h}_i^{(k+1)} m_i^{(k+1)}, \tag{20}$$

where $F_{\text{host}}(m_1^{(k+1)}, \cdots, m_L^{(k+1)})$ is the mean fitness of the population within the EMF approximation.

Parameter estimation

The prevalence landscape of p24 gag

p24 gag, which encodes the HIV capsid, is a highly conserved protein, so considering binarized amino acid sequences as presented above is a reasonable approximation [12].

We follow the methods described in Barton et al. [24] to infer the prevalence landscape of p24 from an alignment of HIV-1 group M subtype B protein sequences (downloaded from [38]). To improve data quality, all sequences with >5% gaps or ambiguous amino acids were removed, and all remaining ambiguous amino acids were imputed by the consensus at that position [24]. Before performing the inference, the sequences were binarized such that the most common (consensus) residue at each position was relabeled 0, and any mutation (or gap) was relabeled 1. To prevent multiple sequences drawn from the same patient introducing biases in the sample mean and second moment, we weighted each sequence by one divided by the number of samples from that patient in the alignment [24]. We fit this distribution of p24 sequences to $P(S) \propto e^{F^p(S)}$, where the prevalence landscape takes the form

$$F^p(S) = \sum_{i=1}^{L} h_i^p s_i + \sum_{i<j} J_{ij}^p s_i s_j, \tag{21}$$

(see [12, 13]), and we inferred the values of $\{h_i^p\}$ and $\{J_{ij}^p\}$ using available techniques [39, 40] (see Fig 3).

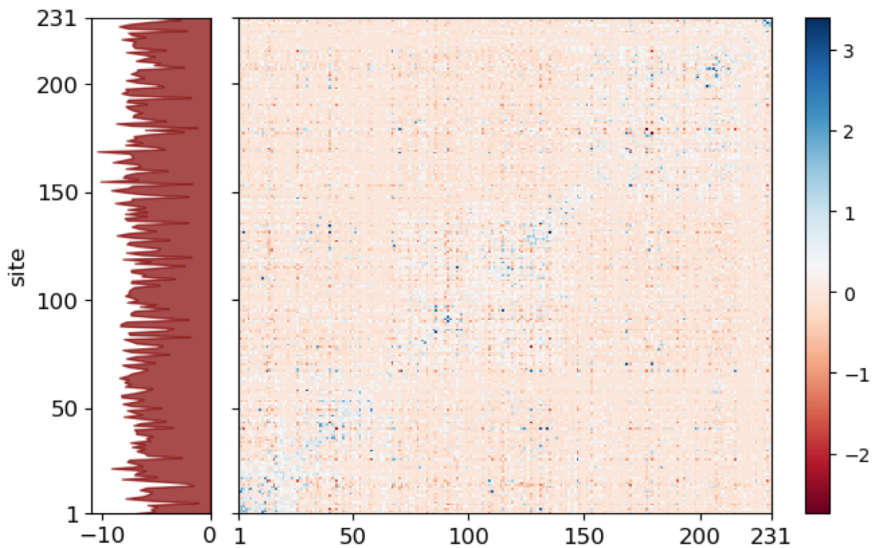

Fig 3. The prevalence landscape of p24. The distribution of p24 protein sequences among HIV-1 group M subtype B strains infecting the global population of hosts (downloaded from [38]) was assumed to take the form $P(S) \propto e^{F^p(S)}$, where $F^p(S) = \sum_i h_i^p s_i + \sum_{i<j} J_{ij}^p s_i s_j$ is the prevalence landscape (see [12, 13]). The values of $\{h_i^p\}$ and $\{J_{ij}^p\}$ were inferred using available techniques [40]. Left panel: Fields h_i^p along the p24 sequence (length $L = 231$). All $h_i^p < 0$ because p24 is highly conserved so all $s_i = 1$ are observed less frequently than $s_i = 0$. Right panel: Color map of couplings J_{ij}^p; positive couplings are in blue and negative couplings are in red.

Estimating the proportionality factor between the prevalence and fitness landscapes, β, and fitness of consensus, F_0

The intrinsic fitness landscape, Eq (1), involves parameters F_0, $\{h_i\}$, and $\{J_{ij}\}$ with units of inverse time, whereas the parameters $\{h_i^p\}$ and $\{J_{ij}^p\}$ in the prevalence landscape, Eq (21), are dimensionless. Shekhar et al. [37] showed that $\{h_i^p\}$ and $\{h_i\}$, and $\{J_{ij}^p\}$ and $\{J_{ij}\}$, are proportionally related in certain regimes, in particular when HIV infects a population of hosts mounting diverse immune responses (which is to some extent satisfied in real life; see e.g. [16]). To find the proportionality constant β, where

$h_i = \beta h_i^p$ and $J_{ij} = \beta J_{ij}^p$, we first turned to the replicative capacity measurements of Mann et al. [13]. Mann et al. found the p24 prevalence landscape to be well correlated with the replicative capacities $RC(S)$ of a number of p24 mutants S growing in cell culture, with a proportionality factor of 0.07 between $RC(S)/RC(S_{\text{NL4-3}})$ and $F^p(S) - F^p(S_{\text{NL4-3}})$, where the reference strain NL4-3 has mutations at sites 120 and 208 within p24 with respect to the HIV-1 group M subtype B consensus used to infer the prevalence landscape. Given that the replicative capacity of the NL4-3 strain is 1.5 day^{-1} [13], this implies a β of $0.07 \times 1.5 = 0.11$ day^{-1}. We found that this value of β produces relative fitness costs, $(F_0 - F_{\text{in}}(S))/F_0$, distributed around 40–50% (cf. Fig 4), which were larger than the site-specific fitness costs inferred from intra-host variation in another study [16] that were (broadly) distributed around 10%. Motivated by [16], we took $\beta = 0.023$ day^{-1}, which gives relative fitness costs of all single and double mutants w.r.t. consensus shown in Fig 4.

Also, using $F^p(S_{\text{NL4-3}}) = -4.13$, $RC(S_{\text{NL4-3}}) = 1.5$ day^{-1}, and this value of β, we found $F_0 - 1.5 = \beta(0 - F^p(S_{\text{NL4-3}})), \;\; \Rightarrow F_0 = 1.6$ day^{-1}. We performed the following back-of-the-envelope check for F_0: around the time of peak viremia—roughly 2 weeks after infection—there are $\sim 10^{10}$ virus particles in a host (because there is a peak viral load of $\sim 10^6$ RNA copies/ml of blood and $\sim$5 liters of blood in a human). Solving $\exp(F \times 14 \text{ days}) = 10^{10}$ gives $F \approx 1.6$ day^{-1}, which is consistent.

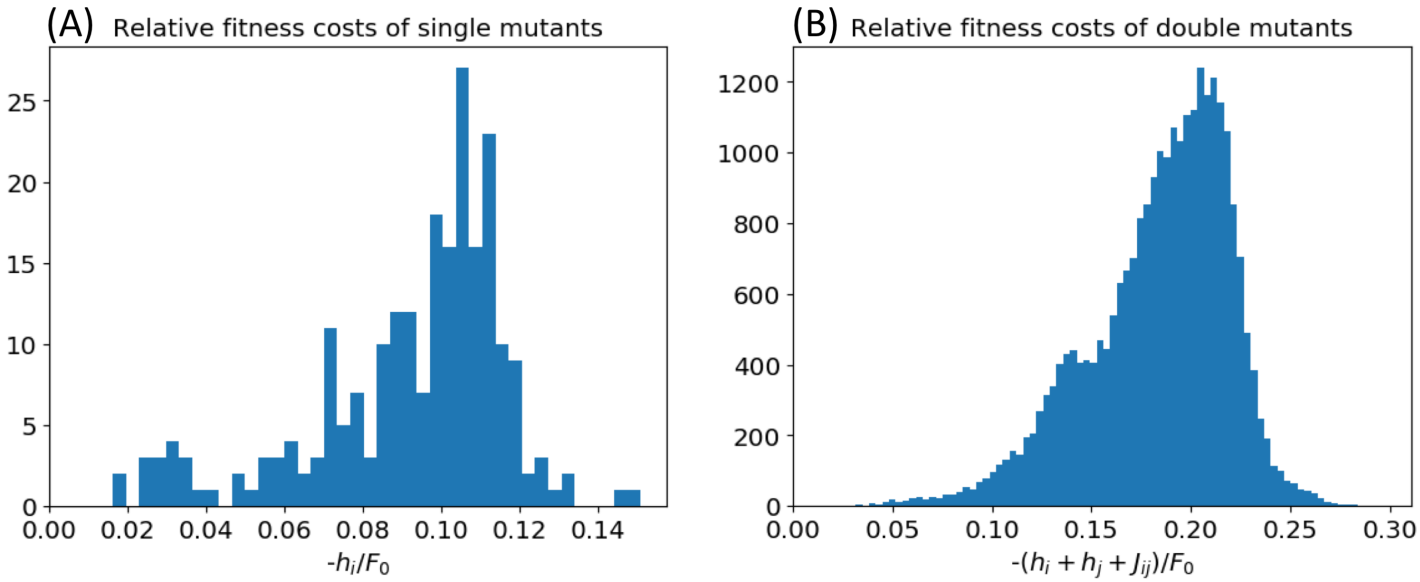

Fig 4. Histograms of relative fitness costs of all single and double mutants according to the inferred fitness landscape, using $\beta = 0.023$ day^{-1}.

Parametrizing the CTL response b_ε

For the kinetics of the CTL response $b_\varepsilon(t)$, we assumed the following Hill-like functional form[4]:

$$b_\varepsilon(t) = \begin{cases} 0, & 0 \le t < t_0 \\ \frac{(t-t_0)^{n_{\text{Hill}}}}{t_{1/2}^{n_{\text{Hill}}} + (t-t_0)^{n_{\text{Hill}}}} b_{\varepsilon,0}, & t \ge t_0, \end{cases} \tag{22}$$

where t_0 is the time taken for the CTL response to be activated post-infection, $b_{\varepsilon,0}$ is the maximum of $b_\varepsilon(t)$, and $t_{1/2}$ and n_{Hill} parametrize the increase of $b_\varepsilon(t)$ to its

[4]We found that HIV dynamics do not depend sensitively on the precise form of $b_\varepsilon(t)$, and in the Discussion we propose an extension to EMF that explicitly accounts for CTL clones.

maximum. As a systemic infection begins around 5–10 days post-infection and CTL responses emerge 2–3 weeks post-infection [1], we took $t_0 = 7$ days, $t_{1/2} = 7$ days, and $n_{\text{Hill}} = 2$. We chose a representative value for the overall magnitude of the CTL response, $b_{\text{tot}} = \sum_\varepsilon b_{\varepsilon,0}$, in order to obtain realistic rates of decline of plasma viral load once CTL responses are activated. Interestingly, we found that for a broad range of b_{tot}, our simulations gave realistic-looking population dynamics curves during the first weeks of infection (see S1 Fig). This is biologically plausible as different untreated hosts presumably have large variations in magnitude of their CTL response, yet significant exponential growth and decline of plasma viral load during acute infection is observed in all hosts. We used a representative value of $b_{\text{tot}} = 6$ day^{-1} for our simulations.

The amino acid substitution rate(s)

The overall HIV nucleotide substitution rate during intra-host infection was found to be 1.2×10^{-5}/nucleotide/day [16], and specifically depends on transition rates between each pair of nucleotides [6, 16]. In the main text, we consider a simplified probability of mutation μ for binary amino acid residues that is three times the nucleotide substitution rate, $\mu = 3.6 \times 10^{-5}$/site/generation. In S1 Appendix, we extend our methods to allow for mutation rates that are site and state-dependent. We also compute the per-generation transition probabilities $\mu_{i,0\to1}$ and $\mu_{i,1\to0}$ for each site i of the founder sequence of the patient studied below, and show that the mutational dynamics are very similar to the main text. Constructing a full amino acid substitution matrix using the codon map requires a multiple-state model, which is beyond the scope of this paper.

Table 1 summarizes all of the parameter values we used.

Table 1. Parameter values used for the application of EMF to the dynamics of p24.

parameter	description	value (units)
F_0	intrinsic fitness of consensus sequence	1.6 day^{-1}
b_{tot}	overall magnitude of host CTL response	6 day^{-1}
L	length of consensus p24 amino acid sequence	231 sites
μ	mutation rate	3.6×10^{-5} site^{-1}day^{-1}
β	proportionality constant relating prevalence and intrinsic fitness landscapes: $h_i = \beta h_i^p$, $J_{ij} = \beta J_{ij}^p$	0.023 day^{-1}
t_0	time delay before ramp-up of host CTL response	7 days
$t_{1/2}$	time to half-maximum of ramp-up	7 days
n_{Hill}	Hill coefficient	2
N_0	initial population size	10

Results

EMF is a high-recombination-rate model of HIV dynamics

In Methods, we derived EMF as an approximation of HIV replication and mutation dynamics, leading to a set of effective fitnesses $\{h_i^{(k)}\}$ and estimates of mutational frequencies $\{m_i^{(k)}\}$ at all HIV residues over time. EMF maps HIV sequence dynamics given a fitness landscape to an approximate model with L *independently* evolving sites experiencing *time-dependent* fitnesses (Fig 2). Frequencies of mutations are simpler to compute using EMF because we no longer have to sum over 2^L terms in Eq (6), which is an advantage for protein sequences of realistic size. While the approximation is

motivated by statistical physics, there is an interesting biological interpretation of the dynamics produced by EMF, which we describe next.

While we have so far only considered HIV replication and mutation, HIV also undergoes recombination during intra-host infection. HIV switches RNA templates during reverse transcription at a rate estimated at 2.8 events/genome/cycle [41]. This may generate novel HIV recombinant sequences only if a target cell is coinfected by multiple HIV strains [42]. The effective HIV recombination rate was estimated from intra-host genetic variation to be 1.4×10^{-5}/nucleotide/day [43], which is of the same order as its mutation rate. Recombination plays an important role during intra-host infection because separate escape mutations recombining onto the same genome may generate a fitter virus, and so models of HIV dynamics with high recombination rates would lead to higher escape rates from multiple epitopes [32, 44].

In the limit of high recombination rate, the probability that mutations at sites i and j are jointly observed on a sequence is simply equal to the product of their frequencies, $m_i m_j$; this is called linkage equilibrium [45]. Because EMF describes sequence (or genotype) dynamics in terms of mutant (or allele) dynamics, it should be viewed as a high-recombination-rate model of HIV dynamics. Indeed, Neher and Shraiman [31] derived an equation analogous to Eq (14) by taking the high-recombination-rate limit of genotype dynamics undergoing selection and recombination. Because EMF is a high-recombination-rate approximation, mutations predicted by this method should occur on a faster timescale than in a model with a finite recombination rate (e.g. [24]), and so the transient dynamics produced by EMF may be inaccurate, but the locations and relative timescales of mutations caused by fitness effects should nevertheless be comparable and consistent. Indeed, in S2 Appendix, we demonstrate consistency of the results presented below with Wright–Fisher simulations including a finite recombination rate, as in Ref [24].

That being said, we argue that EMF is the most efficient method possible to compute HIV dynamics on a fitness landscape with epistasis, as it reduces the problem to non-interacting alleles with background-dependent selection coefficients. If there were no epistasis in the landscape (and the population size very large), this model would exactly capture the replication and mutation dynamics, but epistasis results in making the selection coefficients time-varying (i.e., $h_i \to \tilde{h}_i^{(k)}$) depending on the frequencies of mutations at other sites (see Eqs (14) and (16)).

Application of EMF to predict HIV mutational dynamics within patient CH58

While Barton et al. [24] performed *in silico* HIV dynamics simulations for various HIV proteins infecting a cohort of patients in order to show an association in the locations and relative timescales of escape mutations that emerged, here we instead focus on illustrating the EMF method on a single example, showing how the effective fitnesses and frequencies of mutations yielded by the method could be used to explain various features of the dynamics. Specifically, we study the dynamics of the HIV p24 gag protein infecting patient CH58 [3].

Patient CH58 was a male living in the USA who was infected by an HIV-1 group M subtype B strain [3]. He was not treated with antiretroviral therapy (ART) during the period of study, in accordance with contemporaneous medical protocol. Blood plasma was drawn at several timepoints from which snapshots of the intra-host HIV population were determined by single genome amplification and sequencing [8], and the specificity and magnitude of HIV-specific CTL responses were also mapped by interferon-γ ELISpot assays against overlapping peptides spanning the founder viral sequence [3]. The founder sequence had five mutations w.r.t. the subtype B consensus within p24, and

two epitopes within p24 were targeted by the patient (Fig 5(A), left panel; one epitope in Env and one in Nef were also targeted (not shown)). Mutations away from the founder sequence in each blood sample are listed in Fig 5(A), right panel. Frequencies of mutations within the two p24 epitopes over time are also plotted in Fig 5(B).

(A)

CH58
sites initially non-consensus:
27, 91, 107, 120, 225
p24 epitopes:
sites 15–23
sites 108–117

timepoint (days)	p24 mutations	no. of samples
8	-	6
45	-	7
	110	1
	116	1
85	-	1
	110	2
	116	6
154	110	2
	116	5
239	110	7
	92, 116	1
350	110	2
	11, 110	1
	15, 110	1

(B)

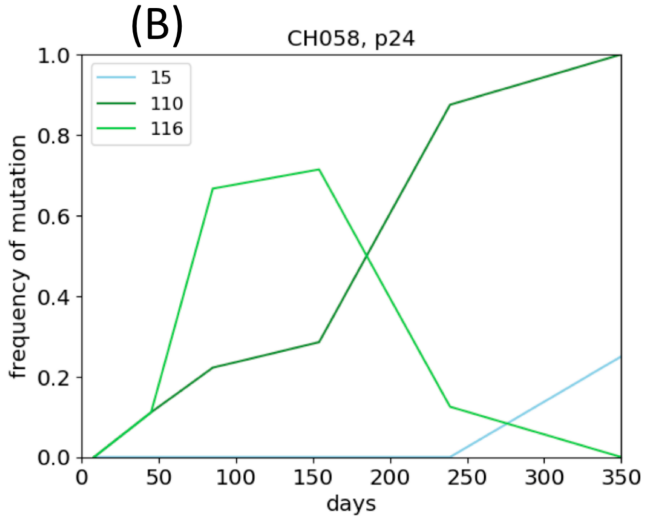

Fig 5. Dynamics of p24 measured in patient CH58 (from [3]). (A) Left panel: Initial binarized p24 sequence and epitopes. Right panel: p24 mutants and sampled frequencies drawn from the patient at multiple timepoints post-infection. Each row represents a distinct sampled sequence. (B) Frequencies of mutations at sites 15, 110 and 116 sampled from patient CH58 over time.

We use the above information, as well as the intrinsic fitness landscape $F_{\text{in}}(S)$ of p24 inferred from an alignment of HIV-1 group M subtype B protein sequences (see Methods), to construct the within-host fitness landscape $F_{\text{host}}(S)$ for patient CH58. In Parameter estimation, we estimated the overall magnitude of the CTL response $b_{\text{tot}} = b_{1,0} + b_{2,0}$, but not the relative values of $b_{1,0}$ and $b_{2,0}$, which are given by the vertical immunodominance hierarchy for the host. We first consider $b_{1,0} = b_{2,0}$, and study the effect of $b_{1,0} \neq b_{2,0}$ later.

EMF outputs coupled dynamics of effective fields and frequencies of mutations

The EMF algorithm starts with the frequencies of mutations of the founder HIV strain(s) (Fig 5(A), left panel). Following Eq (19), we iterate between Eqs (14) and (16), and Eq (18), to obtain the L effective fields $\{\tilde{h}_i^{(k)}\}$ and frequencies of mutations $\{m_i^{(k)}\}$ over time, which are plotted in Fig 6. We find that following the activation of the CTL response (shown in yellow dotted lines), the effective fields at sites within epitopes rise above zero (Fig 6(A)), implying a selective pressure to mutate to $s_i = 1$ at these sites. At longer times, two sites (15 and 110 as shown later) have large $\tilde{h}_i^{(k)} > 0$, while the rest are distributed around a value less than zero.

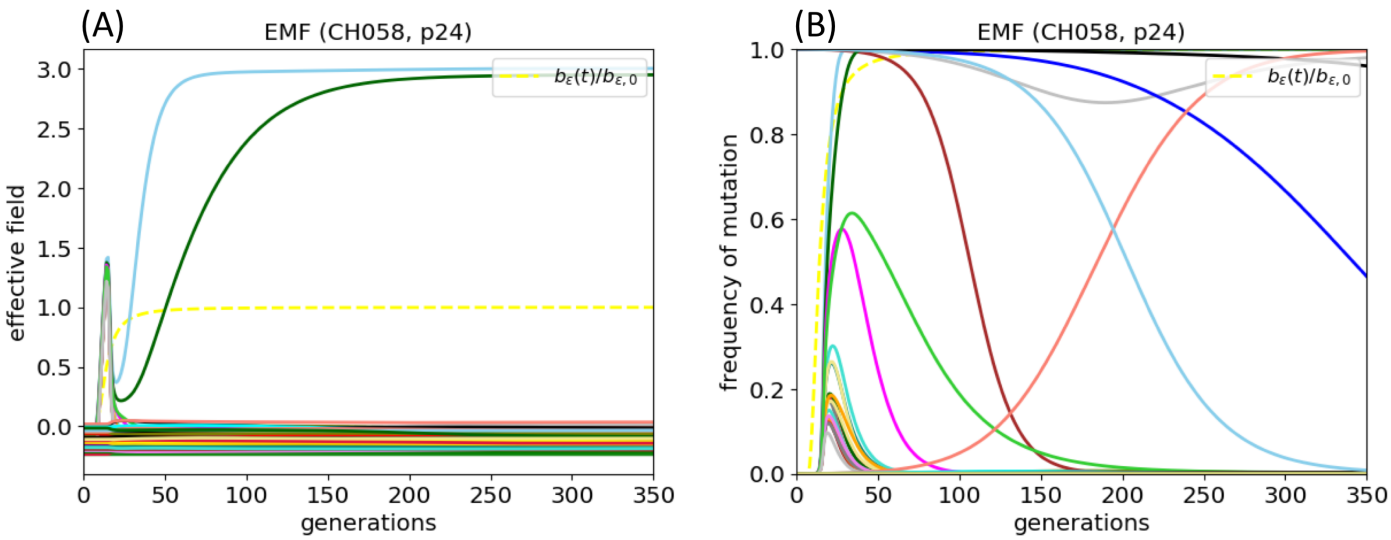

Fig 6. Effective fields and frequencies of mutations produced by EMF. (A) Effective fields $\{\tilde{h}_i^{(k)}\}$, and (B) frequencies of mutations $\{m_i^{(k)}\}$, at each of the L residues of p24 during evolution within patient CH58, as predicted by the EMF method. Also plotted in the yellow dotted lines are $b_\varepsilon(t)/b_{\varepsilon,0}$ (Eq (22)). Note that as we used parameter values in units of day^{-1} (see Table 1), one generation of EMF corresponds to 1 day.

The effective fields dictate the dynamics of mutational frequencies $m_i^{(k)}$ (Fig 6(B)). Positive $\tilde{h}_i^{(k)}$ lead to increasing $m_i^{(k)}$, while negative $\tilde{h}_i^{(k)}$ lead to decreasing $m_i^{(k)}$, with the magnitude of $\tilde{h}_i^{(k)}$ determining the rate of change. In turn, changes in mutational frequencies feed back into the time-varying effective fields through epistatic interactions (terms depending on J_{ij} and b_ε in Eqs (14) and (16)).

EMF combines intrinsic fitness costs and compensatory interactions to predict which escape mutations arise

Focusing on the two epitopes (Fig 7), we find that mutations initially emerge at all of the sites, but eventually one of them fixes while the others decline. This behavior may be seen from the EMF equations: if the frequency of mutation $m_i^{(k)}$ at site i of an epitope is close to 1, the last term of Eq (16) will be small for sites $j \neq i$ within this epitope, but not for site i. There is thus a tendency for other $m_j^{(k)}$ to decline to zero, leaving all of the weight of $b_\varepsilon(k)$ to act on site i and keeping it mutated.

Both intrinsic fitness costs and compensatory interactions determine which escape mutations arise. In the first epitope of patient CH58, site 15 is the least-fitness-cost mutation according to the intrinsic landscape (Fig 7(A) inset) and is the one that fixed.

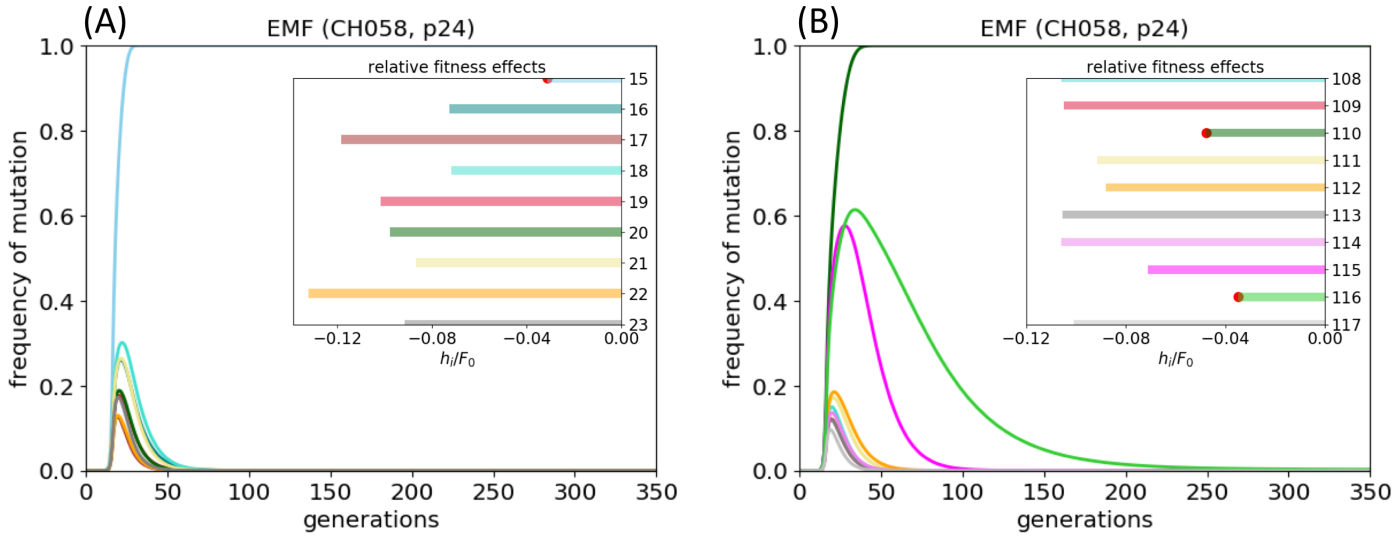

Fig 7. Frequencies of mutations within p24 epitopes targeted by patient CH58 at sites (A) 15–23, and (B) 108–117, as predicted by EMF. Insets show h_i/F_0 at each site; the least-fitness-cost mutations are marked in red.

In the second epitope, site 116 is the least-fitness-cost mutation (Fig 7(B) inset) but site 110 fixed. Indeed, EMF computed the effective field at site 110 to be always larger than the one at site 116 (not shown). Thus, compensatory interactions in the intrinsic landscape induced a different dynamics in patient CH58 from what would be predicted by knowing the fitness costs of mutations alone. This result is consistent with Barton et al. [24] who found this effect for a larger number of patients and proteins, by performing a number of stochastic simulations. Here, we have this information automatically encoded in the effective fitnesses.

In patient CH58, mutations at sites 15, 110 and 116 arose (Fig 5(B)), qualitatively resembling the within-epitope dynamics predicted by EMF. Furthermore, in the patient, mutations at site 116 arose at intermediate times and then disappeared, while mutations at site 110 fixed, qualitatively resembling the predicted dynamics.

CTL immunodominance affects the order and timescale of escape mutations by modifying the relative strengths of effective fields

In patient CH58, mutations at site 15 were only detected on day 350 (Fig 5(B)). This delay might be due to a smaller CTL population targeting the first epitope, and hence a weaker selective pressure for sites 15–23 to mutate. While we had assumed $b_{1,0} = b_{2,0}$ in the EMF computation above, in fact the peak CTL response was measured by ELISpot to be greater by a factor of 4–5 at the second epitope [3]. Different relative magnitudes of CTL responses targeting different epitopes (called the vertical immunodominance) should impact the mutational dynamics that result. How does EMF predict this dynamics to change? We repeated the EMF computation using $b_{1,0} = b_{2,0}/5$, keeping the same overall $b_{\text{tot}} = b_{1,0} + b_{2,0}$. The resulting dynamics are shown in Fig 8(A) (solid lines), predicting that the same escape mutations arise, but unlike the "flat-immunodominance" $b_{1,0} = b_{2,0}$ case (dotted lines), escape at site 15 occurs more slowly than at site 110. This effect would be compounded by a larger time lag t_0 for the first epitope as compared with the second, as was observed in the patient [3] (not shown).

Thus, EMF predicts that CTL immunodominance changes the relative escape rates in the different epitopes, agreeing with the findings of Liu et al. [3]. Also, site 116

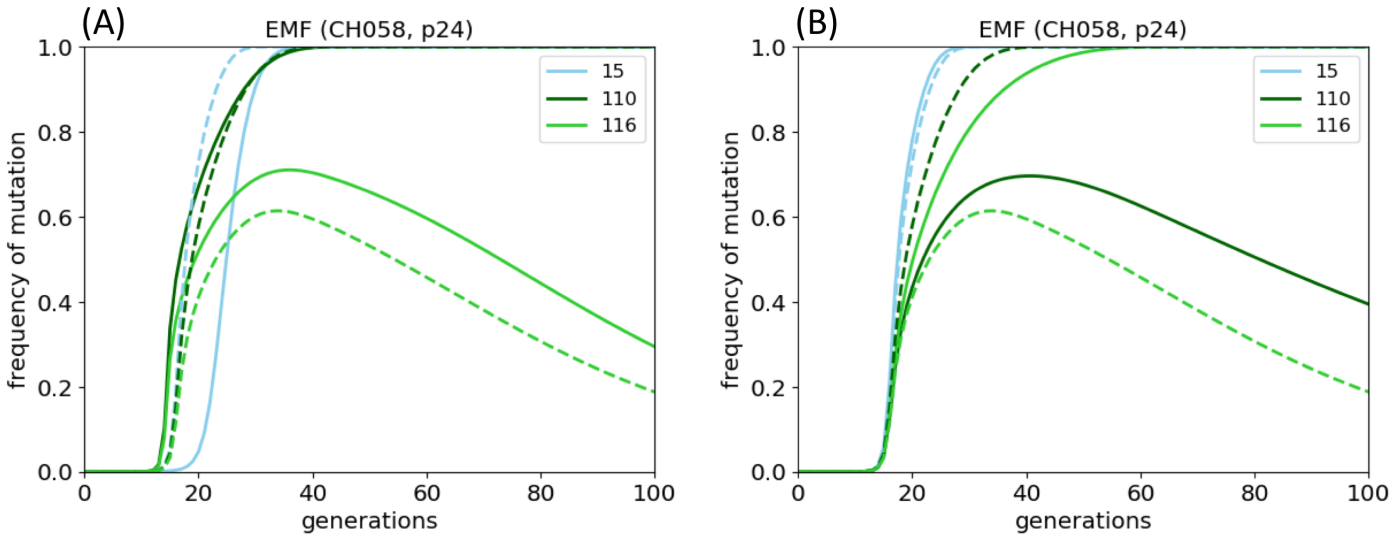

Fig 8. Effect of immunodominance and sequence background on HIV dynamics within patient CH58, as predicted by EMF. (A) Mutational dynamics at sites 15, 110 and 116 when $b_{1,0} = b_{2,0}/5$ (solid lines) and when $b_{1,0} = b_{2,0}$ (dotted lines), for the same $b_{\text{tot}} = b_{1,0} + b_{2,0}$. (B) Mutational dynamics at sites 15, 110 and 116 when the founder sequence is the NL4-3 strain (s_{120} and $s_{208} = 1$). Site 116 mutates instead of 110 according to EMF.

mutations increase to a greater frequency before declining than in the $b_{1,0} = b_{2,0}$ case, suggesting that a sequence background with site 15 mutated leaves site 110 relatively fitter than site 116, as compared with a sequence background with site 15 not mutated (indeed, $J_{15,110} > J_{15,116}$).

A different sequence background may lead to different escape mutations arising because of compensatory interactions

In models of HIV dynamics considering multiple strains that do not include epistatic interactions (e.g., [25–28, 32]), sequence variation outside of the epitope would not affect the location of escape mutations acquired in an epitope. Here, we asked if a different founder sequence would affect the escape mutations that arise in a host mounting the same CTL responses, and computed EMF dynamics starting with, as an example, the NL4-3 strain (which has mutations at sites 120 and 208 w.r.t. $S = (0, \cdots, 0)$) and keeping all other parameters (including epitopes) the same as before. Fig 8(B) shows the result for sites 15, 110 and 116. In contrast with the above results, site 116 acquired the escape mutation in the second epitope. This is indeed the naïve expectation by looking at just the fitness costs in the intrinsic landscape (Fig 7(B) inset). Thus, a knowledge of epistatic interactions in the fitness landscape may enable more accurate predictions of which escape mutations arise in a given host. Furthermore, EMF is a computationally efficient method for making these predictions.

EMF predicts reversions and compensatory mutations occurring over longer timescales

So far, we have only studied the dynamics of escape within epitopes. What does EMF tell us about the mutational dynamics outside of epitopes? Figure 9 shows the effective fields and frequencies of mutations at all sites outside of the two epitopes for patient CH58. The effective fields are distributed around a value less than zero and are mostly negative. Because their magnitudes are close to zero, the dynamics occur over longer

timescales than at sites within epitopes that undergo strong CTL-driven selection. We find that the initially non-consensus sites revert back to consensus over longer timescales, which is consistent with the findings of Refs [10, 11] and especially [9] that found that sites tend to revert to the group or subtype consensus residue throughout the course of intra-patient evolution in all of the patients studied [9]. Within the EMF framework, a distribution in timescales of reversion during intra-host infection is due to a distribution of effective fields that are negative and close to zero.

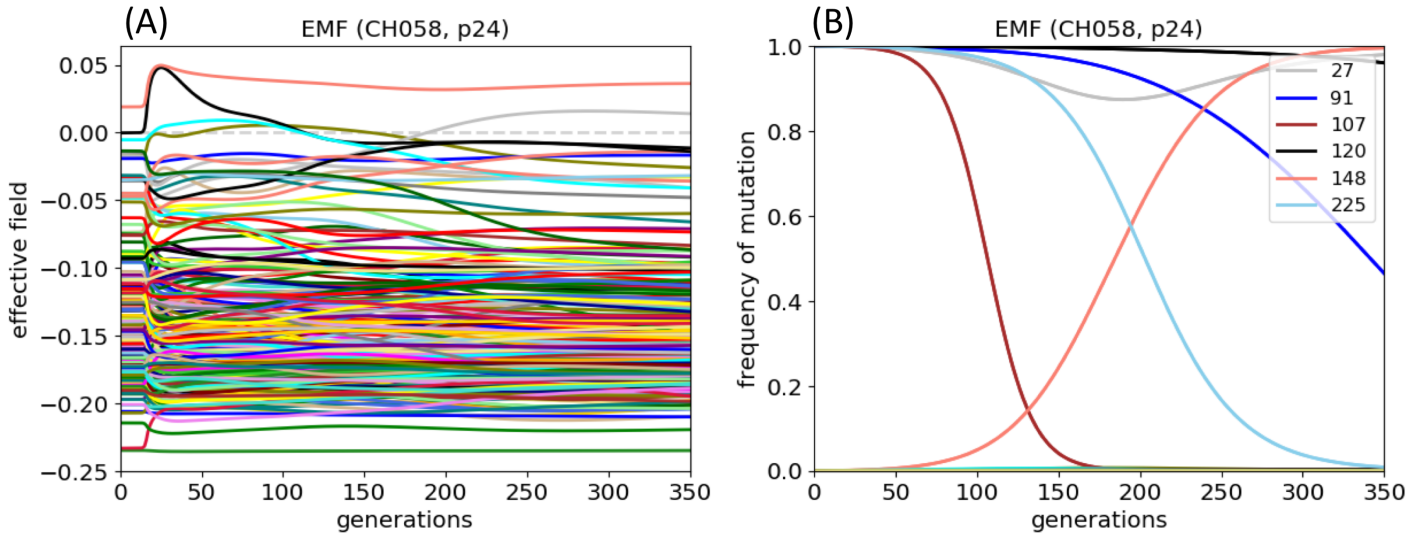

Fig 9. Reversions and compensatory mutations. (A) Effective fields $\{\tilde{h}_i^{(k)}\}$, and (B) frequencies of mutations $\{m_i^{(k)}\}$, at p24 residues outside of epitopes during evolution within patient CH58, as predicted by the EMF method. Compensatory mutations occur at sites whose effective fields rise above zero (here, site 148). Reversions occur at sites initially non-consensus whose effective fields remain negative (here, sites 27, 91, 107, 120, and 225).

Finally, EMF predicts mutations that are neither escape mutations nor reversions but are compensatory due to the (changing) sequence background. For patient CH58, EMF predicts site 148 to mutate ($\tilde{h}_{148} > 0$ at all times in Fig 9(A)). On the other hand, the patient had mutations at sites 92 and 11, which are not predicted by EMF (Fig 5(A)).

We speculate on reasons why mutations predicted by EMF do not occur in the patient, and vice-versa. Above all else, our model is incomplete: Here, we limited ourselves to a binary model, and knowing the fitness landscape for amino acid or nucleotide sequences would likely improve the accuracy of EMF's predictions. The functional form of the fitness landscape used here may also not be complete; there may be higher-order interactions ignored by the model, and the h_i and J_{ij} parameters inferred from prevalence are not quantitatively accurate, as pointed out in other work [14–16]. Furthermore, EMF is an approximate high-recombination-rate model, which neglects crucial features such as genetic hitchhiking. In patient CH58, because the site 92 mutation appeared with the site 116 mutation on day 239 at a low frequency before the site 116 mutation went extinct, it is possible that the site 92 mutation was a hitchhiker mutation, which cannot be accounted for by EMF. For the site 11 mutation, an additional possibility is that it affects antigen processing of the site 15–23 epitope and hence its presentation to CTLs [46, 47], allowing evasion of CTL pressure despite not having mutations within the actual epitope (indeed, the site 11 mutation arose at the same time as the site 15 mutation). Our current model does not account for antigen-processing mutants, although F_{host} can easily be modified to do so.

Finally, mutations observed in the patient but not predicted by EMF might be due

to stochastic fluctuations (i.e., genetic drift) during intra-host evolution, which we will elaborate on in greater detail when we develop a stochastic version of EMF below.

Ultimately, even if we knew the fitness landscape and host CTL responses precisely, the EMF method has ignored numerous details of intra-host HIV infection such as infection of target cells, heterogeneity and spread within host tissues, etc. All these effects presumably influence HIV mutational dynamics as well. In light of this, the qualitative agreement between EMF and intra-patient dynamics in this example is quite encouraging.

Mean fitness and entropy of the intra-host population

The mean fitness and other quantities describing the entire intra-host population are especially simple to compute within the EMF framework, which we show here. The mean fitness at time k is given by

$$\langle F \rangle^{(k)} = F(m_1^{(k)}, \cdots, m_L^{(k)}), \tag{23}$$

which is plotted in Fig 10(A) for $F = F_{\text{host}}$ (blue curve) and $F = F_{\text{in}}$ (orange curve). As host CTLs are activated, the infecting strain becomes strongly selected against, which is represented by the valleys in Fig 10(A). The intra-host population evolves over time to "climb" out of this fitness valley through the generation of mutations and selection of beneficial ones. Thus, the mean fitness first decreases, and then is a strictly nondecreasing function of time.

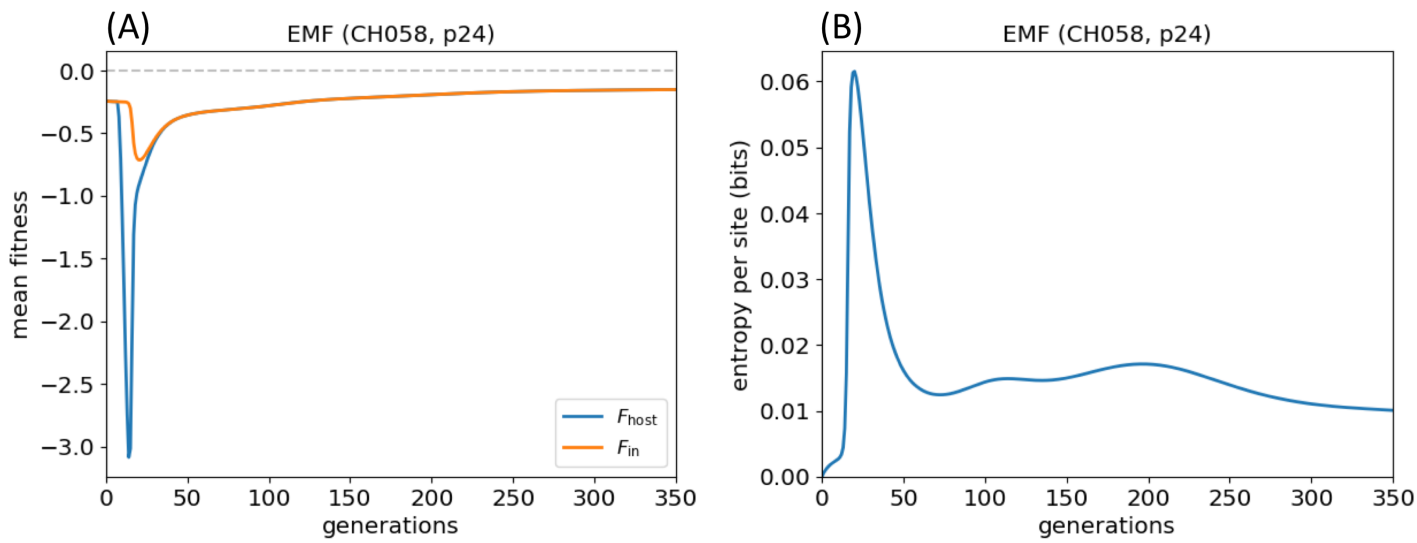

Fig 10. Mean fitness and entropy of the intra-host population over time. (A) Mean fitness $\langle F \rangle^{(k)} - F_0$ of the intra-host HIV population over time within the EMF framework, using $F = F_{\text{host}}$ (blue line) and $F = F_{\text{in}}$ (orange line). The grey dotted line is the fitness of the consensus sequence $(0, \cdots, 0)$. (B) The mean entropy per site over time.

A measure of diversity within the intra-host population is the mean entropy per site,

$$\mathcal{H}^{(k)} = -\frac{1}{L} \sum_{i=1}^{L} \left[m_i^{(k)} \log_2 m_i^{(k)} + (1 - m_i^{(k)}) \log_2 (1 - m_i^{(k)}) \right]. \tag{24}$$

This quantifies how different the frequencies of mutation are, on average, from 0 or 1 ($-m \log_2 m - (1-m) \log_2 (1-m)$ is bounded by 1, has a maximum of 1 at $m = 1/2$, and is 0 at $m = 0, 1$). Figure 10(B) shows that the entropy per site spikes when many

escape mutations arise initially, and declines as they either disappear or fix. Even at long times, the entropy per site remains at a finite value because mutations are continuously generated at all sites but remain at low levels due to being less fit; this is a manifestation of mutation-selection balance [45]. (Indeed, Zanini et al. [16] used mutation-selection balance to infer fitness costs at all sites in the HIV genome not undergoing CTL-driven selection.)

Stochastic population dynamics using the EMF approach

So far, we have focused on the dynamics of mutational frequencies. EMF also outputs the dynamics of the *population size*, since the partition function $Z^{(k+1)}$ (Eq (8)) is precisely the ratio between the population size at time $k+1$ and the population size at time k, given frequencies of mutations $\{m_i^{(k)}\}$ (see Methods). While computing $Z^{(k+1)}$ involves a sum over 2^L terms, the RHS of Eq (11) gives a lower bound to $Z^{(k+1)}$ and is simpler to compute because the approximate model is noninteracting (see Methods).

Also, so far the EMF method always outputs the same dynamics once we specify the initial strain and CTL responses. However, even if we knew these and the intrinsic fitness landscape precisely, the intra-host dynamics is not deterministic because of the many fundamentally stochastic processes involved in a real infection, such as:

- Replication: Identical HIV strains infecting the same target cell type in cell culture have variations in their measured replication rates (see e.g. data in Mann et al. [13]). Stochasticity at the single-cell level would affect the population growth.
- Mutation and genetic drift: Mutations are stochastic events, and "replays" of HIV infection in the same host would lead to different mutations arising at different times. This effect is compounded by the small population size during the early stages of infection (an estimated 80% of mucosally transmitted HIV infections are initiated by a single virus [7]); mutations that happen to arise earlier during infection are more likely to be carried over to when the population size is large.
- Population structure and intra-host spread: The intra-host HIV population infects different tissues within a host which have a variable rate of intermixing between them [48], and some of these subpopulations may have small size. Less fit mutants may fix in some of them because of greater genetic drift. HIV also spreads from its portal of entry and infects various tissues within the host during the early stages of infection, and different viral strains may set up infections in different parts of the host. This process is inherently stochastic.

Computational methods to compute the stochastic dynamics of HIV infection are available, but are largely limited to fixed-population-size models [24, 32]. Zanini and Neher [33] developed a stochastic population dynamics method with a time-dependent population size which is however externally defined and not set by the fitness of the population at each time step. Here, we extend EMF to stochastic population dynamics by taking the population size at generation $k+1$, $N^{(k+1)}$, to be Poisson-distributed [33] with mean $N^{(k)}Z^{(k+1)}$. Importantly, the population dynamics directly depends on which mutations exist and arise in the population at each time. We also draw the frequencies of mutations $\{m_i^{(k+1)}\}$ at time $k+1$ from a binomial distribution with $N^{(k+1)}$ trials and probabilities given by Eq (18), akin to previous stochastic simulation frameworks [24, 32, 33].

EMF population dynamics produces variability in the locations of escape mutations

We apply the stochastic EMF population dynamics method again to the dynamics of p24 gag during infection of patient CH58, assuming the initial population size to be $N_0 = 10$ and keeping all other parameter values the same as before (see Methods). Figures 11(A)–11(C) show the mutational dynamics at sites 15, 110 and 116 for individual stochastic runs (dotted lines), their mean over 200 stochastic runs (dashed lines), and a comparison with the deterministic method (solid lines). Figures 11(D)–11(E) show the mean of 200 stochastic runs (dashed lines) and the deterministic dynamics (solid lines) within the two epitopes.

We find that the dynamics of individual escape events within the epitopes can be very different from the deterministic EMF prediction, as with some (albeit small) probability, other sites may fix. This is why the mean values for sites 15, 110 and 116 are lower for the stochastic case than for the deterministic one, and the averages for other sites within the epitopes are higher. In Figs 11(F)–11(G), we plot the fraction of stochastic runs where each site within each epitope has the largest frequency of mutation at the end of the simulation. For the first epitope, the deterministic dynamics is a very good predictor as site 15 fixes most of the time (other sites may fix if they happen to arise early enough during the simulation). For the second epitope, we find greater variability in the sites that acquire the escape mutation; the deterministic dynamics does predict where the majority of escape mutations arise, and sites 115 and 116 for which the frequency of mutation is relatively large at intermediate times in the deterministic model frequently emerge as the escape mutation in stochastic runs. Thus, predicting for certain which escape mutations arise given accurate knowledge of the HIV fitness landscape and a host's CTL responses may not be possible, but the deterministic version of the dynamics provides good estimates. Note that in our simulations stochasticity arises purely from mutations and genetic drift, and not genetic hitchhiking, because EMF describes the high-recombination-rate limit.

The intra-host population size grows exponentially, then declines due to CTL-mediated killing, consistent with viral load kinetics during acute infection

We compute the dynamics of the population size for patient CH58 using Eq (11). Figure 12 shows the population size changes for the determinstic case (blue line) and 10 stochastic runs (red lines). The population first increases exponentially, then exponentially declines following activation of CTL responses. This is consistent with plasma viral load kinetics in a real infection, as well as experiments that show rapid viral load rebound following CTL depletion that reversed upon replenishment of these cells [5]. Furthermore, this characteristic exponential rise and fall is observed for a range of b_{tot} (see S1 Fig), which is consistent with the fact that viral load kinetics during acute infection occur for diverse hosts who presumably have a range of magnitudes and specificities of CTL responses.

While the exponential increase in viral load and initial decrease after host CTL responses turn on resemble viral load kinetics in a real infection, the intra-host population quickly overcomes host CTL pressure and continues to grow exponentially. In actual patients, this does not happen, for two important reasons:

1. We did not consider an explicit target cell population that declines following infection. Models that include the dynamics of host target cells (via ordinary differential equations) show an approach to a finite fixed point as opposed to unbounded growth [28]. In the Discussion, we explain how target cell dynamics may be incorporated into the EMF approach.

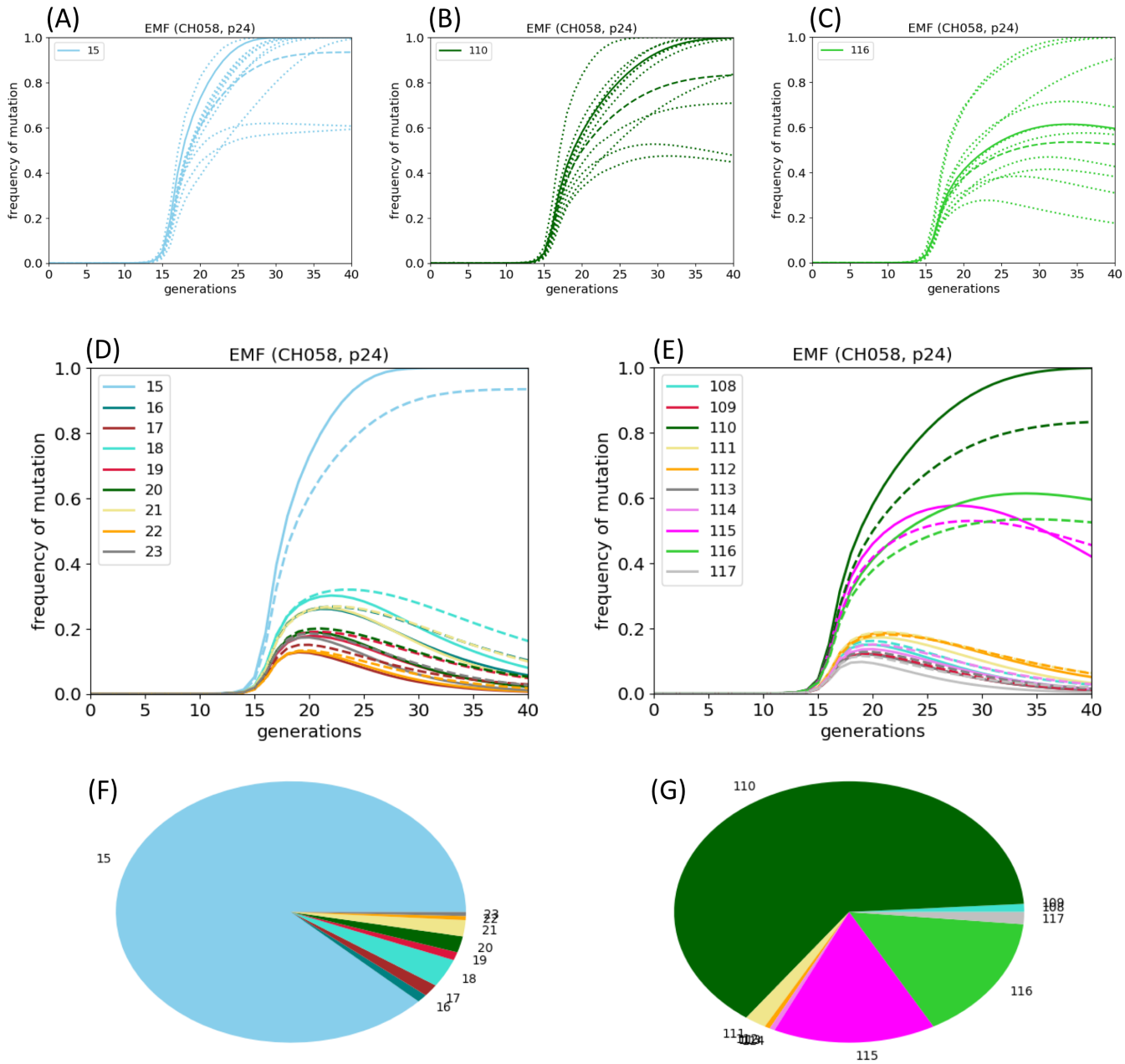

Fig 11. Stochastic dynamics of escape as predicted by EMF population dynamics simulations. (A)–(C) Frequencies of mutations at sites 15, 110 and 116 for the deterministic EMF dynamics (solid lines), 10 representative stochastic runs (dotted lines), and the mean over 200 independent runs (dashed lines). (D)–(E) Frequencies of mutations within the two epitopes for the deterministic EMF dynamics (solid lines) and the mean over 200 independent runs (dashed lines). We do not plot error bars because the dynamics leads to sites having mutational frequencies tending to 0 or 1, so the spread around the mean does not reflect the results of individual stochastic runs. (F)–(G) Fraction of runs where a site has the largest mutant frequency at the end of the run, within each epitope.

2. There are immunodominance shifts in the CTL response. CTLs specific for epitopes from which the HIV population has escaped die out, and new CTL clones specific for other epitopes emerge (see e.g. [25]). Thus, there is continuous CTL

selection and HIV escape during the course of intra-host infection, not captured by the methods as presented here.

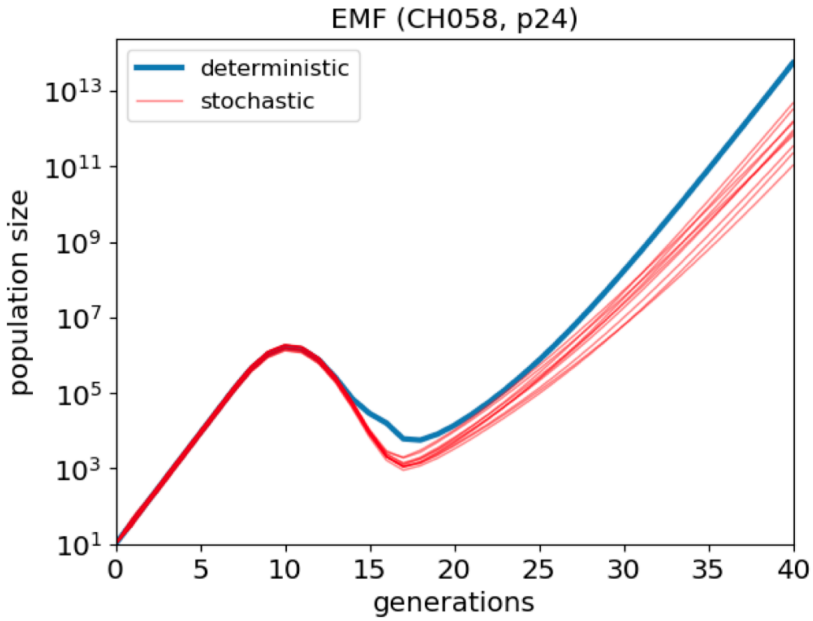

Fig 12. Deterministic and stochastic EMF population dynamics. Dynamics of population size for the deterministic EMF dynamics (blue line) and 10 representative stochastic runs (red lines). The exponential increase and initial decrease in population size resemble plasma viral load kinetics during early acute infection. However, the unbounded exponential growth after viral escape clearly does not reflect a real infection, and is a result of not explicitly considering viral infection and elimination of target cells (see main text).

Discussion

In this paper, we have developed an alternative approach for computing HIV mutational and population dynamics during intra-host infection, which we designate the evolutionary mean-field (EMF) method. EMF takes as input the fitness landscape of an HIV protein, including epistatic interactions, and the locations and strengths of CTL responses in a given host, and outputs "effective fitnesses" describing the tendency of HIV residues to mutate, leading to predictions of the time course of mutations due to these time-dependent fitness effects. EMF makes predictions of mutational dynamics that may otherwise require performing many stochastic simulations [24]. Applying the EMF method to the dynamics of the p24 gag protein evolving within a patient (taken from [3]), we showed how fitness costs and compensatory interactions, vertical immunodominance, and the sequence background, affect the locations and relative timescales of HIV escape mutations, consistent with the work of others [24]. These include cases where knowledge of epistatic interactions improves upon predictions relying simply on fitness costs. We also describe features of longer-term dynamics in terms of these effective fitnesses, specifically reversions which are also qualitatively consistent with other work [9]. We also quantify the mean fitness and entropy of the intra-host population over time within the EMF approximation. Finally, we develop a stochastic population dynamics version of EMF and quantify the variability in the escape mutations that arise. The HIV population sizes resulting from this method show

a characteristic exponential rise and fall during early infection, consistent with observed viral load kinetics, although the unbounded exponential growth following HIV escape is a weakness of the current implementation of EMF, and we suggest below extensions to overcome it.

What have we lost from a high-recombination-rate approximation of HIV dynamics? Unlike a real HIV infection, here escape mutations in separate epitopes quickly recombine, forming variants that escape multiple epitopes; in this sense EMF simulates a "stronger" virus. Also, EMF is unable to account for features such as genetic hitchhiking and clonal interference that do not occur in linkage equilibrium. On the other hand, we expect that the dynamics of mutations due to a fitness landscape are comparable (if we average the dynamics with a finite recombination rate over many runs), and so EMF provides quick predictions that can be checked against simulations implementing a finite recombination rate [24] (see S2 Appendix). Neher and Shraiman [31] obtain equations similar to EMF as the lowest order in an expansion in the inverse recombination rate; it would be interesting to consider the next-order corrections (i.e., quasi-linkage equilibrium) which would extend EMF to account for a finite recombination rate (this would however require $L(L-1)$ extra equations, 5×10^4 for the example that we study).

We have presented EMF as a discretized-time algorithm where each time step represents a replication cycle of HIV (corresponding to $\sim$1 day). This method could be generalized to continuous time by discretizing time more finely and scaling the replication and mutation rates appropriately (viz., if one time step represents $1/M$ generations, then $F_0 \to F_0/M$, $b_{\rm tot} \to b_{\rm tot}/M$, and $\beta \to \beta/M$). This would also lead to more smoothly varying effective fields and frequencies of mutations, which would better justify the "shooting" procedure in Methods. In this limit, EMF effectively describes an exponentially growing population

$$\frac{dN(t)}{dt} = F_{\rm host}(t)N(t), \tag{25}$$

where $F_{\rm host}(t)$ depends on the genetic composition of the population at time t, and CTL pressure is implicit in $F_{\rm host}$. Thus, mutational escape leads to continued exponential growth (see Fig 12). To overcome this limitation, we may consider a target cell population and CTL clones explicitly (much like in other compartmental models [25–28]), in particular:

1. To account for a target cell population and explicit viral infection of target cells, we may extend Eq (25) to a pair of coupled equations:

$$\frac{dT(t)}{dt} = \lambda - d_T T(t) - F_{\rm host}(t)N(t)T(t), \tag{26}$$

$$\frac{dN(t)}{dt} = F_{\rm host}(t)N(t)T(t) - \delta N(t), \tag{27}$$

where λ and d_T are the influx and natural death rates of target cells, and δ is the rate of clearance of viral particles. Effectively, these equations extend two-compartment models of HIV dynamics [28] to have viral infectivity that depends on the genetic composition of the intra-host population. These models produce viral population sizes that approach a finite fixed point [28]; in essence, "fit" viruses do not grow unboundedly if the target cell population is limiting. Within the EMF procedure, we may first compute an intermediate $\bar{N}(t+\delta t)$ using Eq (11), which accounts for population size changes due to changes in frequencies of mutations, and then $T(t+\delta t)$ and $N(t+\delta t)$ using Eqs (26) and (27) (with $\bar{N}$ in place of N on the RHS of these equations).

2. To account for explicit CTL clone dynamics, we may extend Eq (25) to have an equation for each CTL clone specific for each epitope:

$$\frac{dN(t)}{dt} = F_{\text{in}}(t)N(t) - \sum_{\varepsilon} b_{\varepsilon}(t)N_{\varepsilon}(t), \tag{28}$$

$$\frac{db_{\varepsilon}(t)}{dt} = r_{\varepsilon}b_{\varepsilon}(t)(1 - \sum_{\varepsilon} b_{\varepsilon}(t)/b_{\text{tot}})N_{\varepsilon}(t), \tag{29}$$

where r_{ε} is the growth rate of the CTL population specific for epitope ε, $N_{\varepsilon}(t)$ is the HIV subpopulation susceptible to these CTLs, and b_{tot} is now a carrying capacity.

While this paper develops approaches for HIV dynamics *given* a fitness landscape, these methods are presumably limited by the quantitative accuracy of the landscapes used. Here we have estimated the p24 fitness landscape from the prevalence of strains infecting diverse hosts [12, 13], and chosen a value of β such that the distribution of relative fitness costs are around 10% (see Fig 4). While it is known from cell culture experiments that a fraction of p24 mutants are unviable [14], we observed variation at all sites within p24 in the multiple sequence alignment [38], and so the prevalence landscape has a finite value for all binary sequences. Considering multiple states at each site would allow us to model protein or nucleotide sequences that are unviable. More broadly, quantitative estimates of intrinsic fitness from prevalence are limited by several issues, as studied previously [14, 16, 37]; in particular, prevalent HIV mutations are more likely to be HLA-associated than less prevalent ones [16]. Alternatively, measurements of intrinsic fitness from cell culture experiments [14, 15], including epistatic interactions, may eventually lead to more accurately known fitness landscapes of HIV proteins. This may lead to more accurate predictions of the dynamics computed using EMF, and other methods.

Supporting information

S1 Appendix. EMF with state and site-dependent mutation rates.

S2 Appendix. Validation of EMF dynamics results by Wright–Fisher simulations.

S1 Fig. EMF population dynamics produces a characteristic exponential rise and fall of population size during the early stages of infection, for a range of b_{tot}.

Acknowledgments

We thank Arup K. Chakraborty and John P. Barton for insightful discussions and a close reading of the manuscript. We also thank John P. Barton for help with inferring the p24 prevalence landscape. HC was supported by an A*STAR Scholarship and the Ragon Institute of MGH, MIT and Harvard. MK acknowledges support from NSF through grant number DMR-1708280.

References

1. McMichael AJ, Borrow P, Tomaras GD, Goonetilleke N, Haynes BF. The immune response during acute HIV-1 infection: clues for vaccine development. Nat Rev Immunol. 2010;10(1):11–23.

2. Goonetilleke N, Liu MK, Salazar-Gonzalez JF, Ferrari G, Giorgi E, Ganusov VV, et al. The first T cell response to transmitted/founder virus contributes to the control of acute viremia in HIV-1 infection. J Exp Med. 2009;206(6):1253–1272.

3. Liu MK, Hawkins N, Ritchie AJ, Ganusov VV, Whale V, Brackenridge S, et al. Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. J Clin Invest. 2013;123(1):380–393.

4. Abbas AK, Lichtman AH, Pillai S. Cellular and molecular immunology. 9th ed. Elsevier Health Sciences; 2008.

5. Schmitz JE, Kuroda MJ, Santra S, Sasseville VG, Simon MA, Lifton MA, et al. Control of viremia in simian immunodeficiency virus infection by CD8+ lymphocytes. Science. 1999;283(5403):857–860.

6. Mansky LM, Temin HM. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J Virol. 1995;69(8):5087–5094.

7. Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG, et al. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proc Natl Acad Sci USA. 2008;105(21):7552–7557.

8. Salazar-Gonzalez JF, Salazar MG, Keele BF, Learn GH, Giorgi EE, Li H, et al. Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. J Exp Med. 2009;206(6):1273–1289.

9. Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, Albert J, et al. Population genomics of intrapatient HIV-1 evolution. eLife. 2015;4.

10. Friedrich TC, Dodds EJ, Yant LJ, Vojnov L, Rudersdorf R, Cullen C, et al. Reversion of CTL escape-variant immunodeficiency viruses in vivo. Nat Med. 2004;10(3):275–281.

11. Brumme ZL, Brumme CJ, Carlson J, Streeck H, John M, Eichbaum Q, et al. Marked epitope- and allele-specific differences in rates of mutation in human immunodeficiency type 1 (HIV-1) Gag, Pol, and Nef cytotoxic T-lymphocyte epitopes in acute/early HIV-1 infection. J Virol. 2008;82(18):9216–9227.

12. Ferguson AL, Mann JK, Omarjee S, Ndung'u T, Walker BD, Chakraborty AK. Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. Immunity. 2013;38(3):606–617.

13. Mann JK, Barton JP, Ferguson AL, Omarjee S, Walker BD, Chakraborty A, et al. The fitness landscape of HIV-1 gag: advanced modeling approaches and validation of model predictions by in vitro testing. PLoS Comput Biol. 2014;10(8):e1003776.

14. Rihn SJ, Wilson SJ, Loman NJ, Alim M, Bakker SE, Bhella D, et al. Extreme genetic fragility of the HIV-1 capsid. PLoS Pathog. 2013;9(6):e1003461.

15. Rihn SJ, Hughes J, Wilson SJ, Bieniasz PD. Uneven genetic robustness of HIV-1 integrase. J Virol. 2015;89(1):552–567.

16. Zanini F, Puller V, Brodin J, Albert J, Neher RA. In vivo mutation rates and the landscape of fitness costs of HIV-1. Virus Evol. 2017;3(1):vex003.

17. Crawford H, Prado JG, Leslie A, Hue S, Honeyborne I, Reddy S, et al. Compensatory mutation partially restores fitness and delays reversion of escape mutation within the immunodominant HLA-B*5703-restricted Gag epitope in chronic human immunodeficiency virus type 1 infection. J Virol. 2007;81(15):8346–8351.

18. Schneidewind A, Brockman MA, Yang R, Adam RI, Li B, Le Gall S, et al. Escape from the dominant HLA-B27-restricted cytotoxic T-lymphocyte response in Gag is associated with a dramatic reduction in human immunodeficiency virus type 1 replication. J Virol. 2007;81(22):12382–12393.

19. Brockman MA, Schneidewind A, Lahaie M, Schmidt A, Miura T, Desouza I, et al. Escape and compensation from early HLA-B57-mediated cytotoxic T-lymphocyte pressure on human immunodeficiency virus type 1 Gag alter capsid interactions with cyclophilin A. J Virol. 2007;81(22):12608–12618.

20. Brockman MA, Brumme ZL, Brumme CJ, Miura T, Sela J, Rosato PC, et al. Early selection in Gag by protective HLA alleles contributes to reduced HIV-1 replication capacity that may be largely compensated for in chronic infection. J Virol. 2010;84(22):11937–11949.

21. Dahirel V, Shekhar K, Pereyra F, Miura T, Artyomov M, Talsania S, et al. Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. Proc Natl Acad Sci USA. 2011;108(28):11530–11535.

22. Poon AF, Swenson LC, Dong WW, Deng W, Kosakovsky Pond SL, Brumme ZL, et al. Phylogenetic analysis of population-based and deep sequencing data to identify coevolving sites in the nef gene of HIV-1. Mol Biol Evol. 2010;27(4):819–832.

23. Louie RHY, Kaczorowski KJ, Barton JP, Chakraborty AK, McKay MR. Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies. Proc Natl Acad Sci USA. 2018;115(4):E564–E573.

24. Barton JP, Goonetilleke N, Butler TC, Walker BD, McMichael AJ, Chakraborty AK. Relative rate and location of intra-host HIV evolution to evade cellular immunity are predictable. Nat Comm. 2016;7:11660.

25. Althaus CL, De Boer RJ. Dynamics of immune escape during HIV/SIV infection. PLoS Comput Biol. 2008;4(7):e1000103.

26. Ganusov VV, Goonetilleke N, Liu MK, Ferrari G, Shaw GM, McMichael AJ, et al. Fitness costs and diversity of the cytotoxic T lymphocyte (CTL) response determine the rate of CTL escape during acute and chronic phases of HIV infection. J Virol. 2011;85(20):10518–10528.

27. Ganusov VV, Neher RA, Perelson AS. Mathematical modeling of escape of HIV from cytotoxic T lymphocyte responses. J Stat Mech. 2013;2013:P01010.

28. Perelson AS, Ribeiro RM. Modeling the within-host dynamics of HIV infection. BMC Biol. 2013;11:96.

29. Leuthäusser I. An exact correspondence between Eigen's evolution model and a two-dimensional Ising system. J Chem Phys. 1986;84(3):1884–1885.

30. Leuthäusser I. Statistical mechanics of Eigen's evolution model. J Stat Phys. 1987;48(1-2):343–360.

31. Neher RA, Shraiman BI. Statistical genetics and evolution of quantitative traits. Rev Mod Phys. 2011;83(4):1283–1300.

32. Batorsky R, Kearney MF, Palmer SE, Maldarelli F, Rouzine IM, Coffin JM. Estimate of effective recombination rate and average selection coefficient for HIV in chronic infection. Proc Natl Acad Sci USA. 2011;108(14):5661–5666.

33. Zanini F, Neher RA. FFPopSim: an efficient forward simulation package for the evolution of large populations. Bioinformatics. 2012;28(24):3332–3333.

34. Markowitz M, Louie M, Hurley A, Sun E, Di Mascio M, Perelson AS, et al. A novel antiviral intervention results in more accurate assessment of human immunodeficiency virus type 1 replication dynamics and T-cell decay in vivo. J Virol. 2003;77(8):5037–5038.

35. Sella G, Hirsh AE. The application of statistical physics to evolutionary biology. Proc Natl Acad Sci USA. 2005;102(27):9541–9546.

36. MacKay DJC. Information theory, inference and learning algorithms. Cambridge University Press; 2003.

37. Shekhar K, Ruberman CF, Ferguson AL, Barton JP, Kardar M, Chakraborty AK. Spin models inferred from patient-derived viral sequence data faithfully describe HIV fitness landscapes. Phys Rev E Stat Nonlin Soft Matter Phys. 2013;88(6):062705.

38. HIV sequence database search interface;. Available from: `https://www.hiv.lanl.gov/components/sequence/HIV/search/search.html` [cited November 2, 2018].

39. Cocco S, Monasson R. Adaptive cluster expansion for inferring Boltzmann machines with noisy data. Phys Rev Lett. 2011;106(9):090601.

40. Barton JP, De Leonardis E, Coucke A, Cocco S. ACE: adaptive cluster expansion for maximum entropy graphical model inference. Bioinformatics. 2016;32(20):3089–3097.

41. Zhuang J, Jetzt AE, Sun G, Yu H, Klarmann G, Ron Y, et al. Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. J Virol. 2002;76(22):11273–11282.

42. Levy DN, Aldrovandi GM, Kutsch O, Shaw GM. Dynamics of HIV-1 recombination in its natural target cells. Proc Natl Acad Sci USA. 2004;101(12):4204–4209.

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

44. Nora T, Charpentier C, Tenaillon O, Hoede C, Clavel F, Hance AJ. Contribution of recombination to the evolution of human immunodeficiency viruses expressing resistance to antiretroviral treatment. J Virol. 2007;81(14):7620–7628.

45. Gillespie JH. Population genetics: a concise guide. 2nd ed. Johns Hopkins University Press; 2004.

46. Draenert R, Le Gall S, Pfafferott KJ, Leslie AJ, Chetty P, Brander C, et al. Immune selection for altered antigen processing leads to cytotoxic T lymphocyte escape in chronic HIV-1 infection. J Exp Med. 2004;199(7):905–915.

47. Jones NA, Wei X, Flower DR, Wong M, Michor F, Saag MS, et al. Determinants of human immunodeficiency virus type 1 escape from the primary CD8+ cytotoxic T lymphocyte response. J Exp Med. 2004;200(10):1243–1256.

48. Feder AF, Kline C, Polacino P, Cottrell M, Kashuba ADM, Keele BF, et al. A spatio-temporal assessment of simian/human immunodeficiency virus (SHIV) evolution reveals a highly dynamic process within the host. PLoS Pathog. 2017;13(5):e1006358.