Table of Contents

Table S1. Parameters of the Basic Model of Initial and Systemic Infection, Related to Figure 1

Notation	Definition	Units	Value	Reference
R_0^{muc}	Reproduction ratio in initial infection	dimensionless	= 0.1 Fig. 2A & Fig. S2A-B =0.25 Fig. S2F	(Li et al., 2005; Miller et al., 2005)
R_0^{LT}	Reproduction ratio in systemic infection	dimensionless	=10 Fig 2 =15* Fig 4-5 & S3-S4	(Nowak et al., 1997)
b_{muc}	Linear replenishment rate of target cells in initial infection	[cells/day]	$R_0^{muc} = b_{muc}\, nk/cd_T$	**
b_{LT}	Linear replenishment rate of target cells in systemic infection	[cells/day]	$2 \cdot 10^{11}\, d_T$	***
k	Virus infectivity parameter	[1/day/RNA copy]		
n	Number of virions from an infected cell	[RNA copy/cell]	$R_0^{LT} = \frac{b_{LT} nk}{cd_T}$	**
c	Virion clearance rate	[1/day]		
d_T	Death rate of target cells	[1/day]	0.1	(Nowak et al., 1997) (Stafford et al., 2000)
d_I	Death rate of virus-producing cells	[1/day]	1.0	(Klatt et al., 2010)
d_L	Death rate of latent cells	[1/day]	10^{-3}-10^{-4}	(Finzi et al., 1997; Sedaghat et al., 2008)
r	Activation rate of latent cells	[1/day]	init. infec.: $6 \cdot 10^{-3}$ systemic infec.: $r(0) = 10^{-3}$-10^{-4} $r(E)$ = Eq. [S10]	(Finzi et al., 1997; Sedaghat et al., 2008)
p_{lat} or p	Probability of latency	*dimensionless*	Free parameter	****
f_{act}	Fraction of systemic infections via non-latent routes at $p = 0.5$	*dimensionless*	Free parameter only in Fig. 2E, S2H	N/A
$f_{nonlatent}$	Fraction of systemic infections via non-latent routes at $p = p_{opt}$	*dimensionless*	Eq. [S13]	Function of p, f_{act}

* When an immune response is incorporated, R_0^{LT} is increased by a factor of 1.5 to account for the viral eclipse phase, which slows down the predicted viral expansion (Klenerman et al., 1996b; Sergeev et al., 2010b).

** Parameters b_{muc}, k, c, n are set at arbitrary values constrained to match the relevant value of R_0

*** b_{LT} corresponds to the total CD4 cell count of $2 \cdot 10^{11}$ before infection (Murphy, 2011).
**** p_{opt} only depends on R_0^{LT} (Figures 2C-D)

Table S2. Additional Parameters of the Immune-Response Model, Related to Figure 3

Notation	Definition	Units	Value	Reference
a	Maximum proliferation rate of CTLs	[1/day]	a-d_E = 0.5	(De Boer et al., 2003)
d_E	CTL death rate	[1/day]	0.75	Data fitting*
d_{IE}	Inverse length of eclipse phase	[1/day]	1.4/day	(Brandin et al., 2006; Markowitz et al., 2003)
E_0	Number of CTLs that halves the infected cell lifetime	cells	$2 \cdot 10^8$	Data fitting*
I_{av}	CTL avidity threshold in the number of infected cells	cells	10^8	Data fitting*
$E_N(0)$	Initial naive CTL number	cells	$2 \cdot 10^6$	(Murphy, 2011)
E_{0L}	Number of CTLs that decreases p_{lat} by 50%	cells	$4 \cdot 10^6$	Data fitting*

* Out of the seven parameters, three (a-d_E, $E_N(0)$, d_{IE}) are fixed and cited from the literature, and four (d_E, E_0, E_{0L}, I_{av}) are fitted to match the four experimental plateaus measured in patients. As shown in the main text (Figure 4), E_0, I_{av}, and E_{0L} are adjusted to fit the measured steady state levels: $E = 10^9$ cells (Ogg et al., 1998) (Turnbull et al., 2009), $I = 10^8$ cells (Haase, 1999), and $L = 10^6$ cells (Chun et al., 1997). The 4th fitted parameter d_E is adjusted to fit L under ART: $L = 10^5$ cells (Finzi et al., 1997). Finally, total cell counts are assumed to be $T(0) = b/d_T = 2 \cdot 10^{11}$ for both $CD8^+$ T and $CD4^+$ T cells (Murphy, 2011).

EXTENDED EXPERIMENTAL PROCEDURES

Notation remark: For brevity, below we use p and p_{lat} interchangeably.

A. One-compartment model ($R_0^{muc} < 1$)

"Chemical reaction" representation

We first describe the dynamics of initial mucosal infection by incorporating latently infected cells into an established stochastic model of early infection (Pearson et al., 2011):

$$\begin{aligned} &\xrightarrow{b} T \\ T &\xrightarrow{d_T} \emptyset \\ T + V &\xrightarrow{(1-p)\mathrm{k}} I \\ T + V &\xrightarrow{p\,k} L \\ L &\xrightarrow{r} I \\ I &\xrightarrow{n\,d_I} I + V \\ I &\xrightarrow{d_I} \emptyset \\ V &\xrightarrow{c} \emptyset \\ L &\xrightarrow{d_L} \emptyset \end{aligned} \tag{S1}$$

In this simplified model, V denotes the number of free virus particles, I denotes actively infected (virus-producing) cells, and L represents the number of latently infected cells. Cells become infected by virus at a rate k and become latently infected with probability p_{lat} and reactivate from latency at a rate r. Infected cells produce n viral particles (burst size) and decay at rate d_I, while virus decays at the rate c. Target cell dynamics (T) can either be explicitly considered (as above) or represented as a constant $T = b/d_T$, when viral loads are small.

Stochastic Gillespie simulation (Fig S2A, B)

The stochastic initial infection model (Eqs S1) was simulated using the Monte-Carlo 'Gillespie' algorithm (Gillespie, 1977). These simulations were implemented via the xSSA package in Mathematica™ with the parameters: $R_0^{muc} = 0.1$, $c = 5$/day (Perelson et al., 1996), $n = 500$ (Zhang et al., 1999), and $k = (cd_T/bn)R_0^{muc}$. Parameter values are summarized in Table S1, with simulation results shown in Figs. S2 A,B.

Deterministic approximation (ordinary differential equations) (Fig 2A, B)

We next simulated the dynamics of initial and systemic infection using a deterministic model that incorporates proviral latency (Rong and Perelson, 2009a, b; Sedaghat et al., 2007; Sedaghat et al., 2008). This deterministic model represents a mean-field approximation to the chemical reaction model in Eqs S1 and is represented by the following coupled system of nonlinear ordinary differential equations:

$$
\begin{aligned}
\text{Uninfected 'target' cells} \quad & \frac{dT}{dt} = \underbrace{b}_{\text{replenishment}} - \underbrace{d_T T}_{\text{natural death}} - \underbrace{kVT}_{\text{infection}} \\
\text{Actively infected cells} \quad & \frac{dI}{dt} = \underbrace{(1-p_{lat})kVT}_{\text{active infection}} - \underbrace{d_I I}_{\text{death}} + \underbrace{rL}_{\text{reactivation}} \\
\text{Latently infected cells} \quad & \frac{dL}{dt} = \underbrace{p_{lat} kVT}_{\text{latent infection}} - \underbrace{d_L L}_{\text{death}} - \underbrace{rL}_{\text{reactivation}} \\
\text{Virus} \quad & \frac{dV}{dt} = \underbrace{nd_I I}_{\text{production}} - \underbrace{cV}_{\text{clearance}}
\end{aligned}
\tag{S2}
$$

Here, uninfected 'target' cells (T) are produced at rate b, decay at rate d_T, and can be infected by virus particles (V) at rate k. Target cells are considered because this deterministic model is capable of describing the full span of systemic HIV infection and accounts for the depletion of uninfected target cells by the virus. Upon virus infection, target cells become either latently infected cells (L) with probability p_{lat} or become actively infected virus-producing cells (I) with probability 1 - p_{lat}. Latently infected cells decay at the relatively slow rate d_L and reactivate into actively infected (virus-producing) cells at rate r with actively infected cells decaying at a rate d_I. Actively infected cells produce virus at the per-capita 'burst-size' n as they decay. All cell types and viral particles have finite lifetimes. Target cells, actively infected cells, and latently infected cells all decay at relatively slow rates d_T, d_I, d_L, respectively, while viral particles decay at a relatively fast rate c. Parameter values (Table S1) were derived from measurements in human patients and are the same for initial and systemic infection periods, with one exception: the replenishment rate of uninfected 'target' cells (b) is assumed to be much lower for the initial infection to account for the lower basic reproduction number $R_0 = kbn/(cd_T)$ measured during the first 5 days of initial infection (Haase, 2011; Li et al., 2005; Miller et al., 2005; Zhang et al., 1999). We solve Eqs S2 numerically using the ODE15s (stiff) solver in MATLAB™.

Infection in the mucosa as a branching process

To obtain the probability of extinction during initial mucosal infection as a function of R_0 and the infected cell inoculum (I_0), we developed an analytic branching process approximation—since individual Monte Carlo runs are inefficient for probing extinction probabilities across a broad range of (R_0, I_0) parameter space.

Here we derive the branching process approximation. Given that the initial inoculum of infected cells in the mucosa is modest, $I_{\text{inoculum}} < 100$ infectious units, the reactivation of latent cells can be neglected (since $r < 10^{-3}$, Table S1). Thus, the number of actively infected cells in the mucosa, $I_{\text{muc}}(t)$, is decoupled from the other two variables: L_{LT} and I_{LT} [i.e., these two are found after $I_{\text{muc}}(t)$ is calculated]. Since HIV virions are short-lived, $c >> d_{\text{I}}$ and d_T (Perelson et al., 1996), their concentration is proportional, as a function of time, to the infected cell number. Excluding viral load as a free variable, we consider the random dynamics of infected cells number $I_{\text{muc}}(t)$ as a branching process (Grimmett and Stirzaker, 2001). At $p = 0$, actively infected cells die at rate d_{I} and leave new infected cells at rate d_{I} R_0^{muc}, where $R_0^{\text{muc}} = k\, b_{\text{muc}}\, n/(cd)$ is the basic reproduction number (number of new cells per old cell). The total number of target cells is assumed constant, because viral load is small, and target cells are not depleted. At any $p > 0$, we have

$$\frac{\partial P(I_{muc}\,,t)}{\partial t} = (I_{muc}-1)(1-p\,)R_0^{muc}P(I_{muc}-1\,,t) \\ +(I_{muc}+1)P(I_{muc}+1\,,t) - I_{muc}[1+(1-p\,)R_0^{muc}]P(I_{muc}\,,t)$$

where $P(I_{muc}\,,t)$ is the probability of having I_{muc} = 0, 1, 2, 3, … actively infected cells at time *t*. The equation was solved numerically using Matlab™ with the initial condition:

$$P(I_{muc},0) = \frac{\big((1-p)I_0\,\big)^{I_{muc}}}{I_{muc}!}\,e^{-(1-p)I_0}$$

We can also add latent cells to this process neglecting their reactivation (see above). We introduce the probability $G(L\,,t)$ of having L = 0, 1, 2, 3,… latent cells at time *t* averaged over all possible random trajectories $P(I_{muc}\,,t)$. The averaging is allowed because the random process for active cells does not depend on that for latent cells and can be calculated first (see previous paragraph). We get:

$$\frac{\partial G(L\,,t)}{\partial t} = [G(L-1\,,t) - G(L,t)] \sum_{I_{muc}=1}^{\infty} \big(pI_{muc}\,P(I_{muc}\,,t)\big)$$

The initial condition is:

$$G(L,0) = \tfrac{(p\,I_0\,)^L}{L!}\,e^{-p\,I_0}\,.$$

The probability of having zero latent cells at time *t*, $G(0,t)$, is readily derived from these equations as:

$$G(0,t) = e^{-p\,I_0}e^{-p\int_0^t dt\prime\,\Sigma_{I_{muc}=1}^{\infty} I_{muc}\,P(I_{muc}\,,t)}$$

Extinction in the mucosa: branching process (Fig S2C, D)

The probability of active-cell extinction at time *t* [condition $I_{muc}(t) = 0$] is given by $P(0,t)$ calculated numerically as explained in the previous subsection. The probability of complete extinction at time *t*, including latent cells [condition $I_{muc}(t) = 0$ & $L(t) = 0$], is given by the product $P(0,t)G(0,t)$. This extinction probability is plotted in Figs S2C, D.

Extinction and surviving cells in the mucosa: analytic Wright-Fisher process (Fig S2E)

The simplest method to calculate the average number of latent cells surviving mucosal infection (Fig S2) is to use a discrete-time Markov process, which represents a generalized version of Wright-Fisher process in which the number of cells is not fixed, as in the classical version (Nielsen and Slatkin, 2013), but changes with time. Assuming a modest inoculum in mucosa, $I_0 < 100$ infectious units, the reactivation of latent cells in mucosa is again neglected (rate $r < 10^{-3}$, Table S1). The dynamics of actively infected and latently infected cells in the mucosa are then derived from the recursive relations:

$$\bar{I}(0) = (1-p)I_0$$
$$\bar{I}(t) = R_0^{\mathrm{muc}}\,(1-p)\bar{I}(t-1),\;\; t = 1,2,3\,\ldots$$

$$\bar{L}'(t) = R_0^{\mathrm{muc}}\, p\bar{I}(t-1)$$

$$\bar{L}(t) = pI_0 + \sum_{i=1}^{t} \bar{L}'(i)$$

where the bar denotes averaging over realizations, and $\bar{L}'(t)$ is the fraction of latent cells born in generation *t*. The initial latent cells, $\bar{L}(0) = pI_{\mathrm{inoculum}}$, come directly from the infecting virus. Iterating these relations, one gets

$$\bar{I}(t) = I_0(1-p)a^t$$

$$\bar{L}(t) = I_0\, p\, \left[1 + a\frac{1-a^{t-1}}{1-a}\right]$$

$$a \equiv R_0^{\mathrm{muc}}(1-p)$$

From these expressions we obtain the number of surviving infected cells $\bar{I}(t) + \bar{L}(t)$ (Fig S2E). In the long run, the average number of cells that survive to start systemic infection in the lymphoid tissue is the output of a geometric series:

$$L_{init}^{R_0>1} = I_0 \frac{p}{1-a} \tag{S3}$$

If R_0^{muc}<<1, then *a* is small. As a result, we obtain the main text assumption (and the simulation result of Figure S2E) that $L_{init}^{R_0>1} \approx I_0 p$.

An intuitive way to understand these equations is to note that in the first generation of mucosal infection, the number of latently infected cells produced is $p_{\mathrm{lat}} I_0$. In each subsequent generation, the production of latently infected cells decreases geometrically by the multiplicative factor $(1\text{-}p_{\mathrm{lat}})R_0^{\mathrm{muc}}$, because new latently infected cells only emerge from actively infected cells that managed to replicate in the previous generation. Given R_0^{muc}<<1, subsequent mucosal generations contribute increasingly small amounts to the infected cell population. For example, when $R_0^{\mathrm{muc}} = 0.25$ and p_{lat}=0.5, 86% of latently infected cells are generated during the first mucosal generation. Thus, the number of infected cells that survive mucosal infection is $\approx p_{\mathrm{lat}} I_0$.

Also, we can estimate the active infection extinction time, $t_{\mathrm{act\_extinct}}$, from the condition $\bar{I}\left(t_{\mathrm{act\_extinct}}\right) = I_{\mathrm{cutoff}}$, where the extinction threshold, $I_{\mathrm{cutoff}} \sim 0.3$ cells, is adjusted to match the Gillespie simulation results (Figs. S2A,B):

$$t_{\mathrm{act\ extinct}} = \frac{\log\left(\frac{I_0}{I_{\mathrm{cutoff}}}\right)}{\left|\log\left[(1-p)R_0^{\mathrm{muc}}\right]\right|}$$

We can also obtain the time when the total number of infected cells is extinct from $\bar{I}(t) + \bar{L}(t) = I_{\mathrm{cutoff}}$.

$$t_{\mathrm{all\ extinct}} = \frac{\log\left(\frac{\frac{I_{\mathrm{cutoff}}}{I_0} - \frac{p}{1-a}}{1-\frac{p(2-a)}{1-a}}\right)}{|\log a|}$$

If p is sufficiently large, this time is infinite, because the number of latent cells surviving infection exceeds I_{cutoff}.

B. Two-compartment model ($R_0^{\text{muc}} < 1$ and $R_0^{\text{LT}} > 1$)

Coupled compartments: Wright-Fisher simulation (Fig S2F-I)

To illustrate the time delay between viral inoculation and systemic viral expansion in the lymphoid tissue (LT), we simulated a stochastic coupled process with explicit cell transfer between the two compartments. In contrast to the simulations in Fig S2A-B, which consider only initial mucosal infection, here we include both initial infection in the mucosa and systemic infection across the lymphoid tissue in a two-compartment model.

In particular, the two-compartment model tracks the first 2 to 3 weeks post-transmission, when the virus levels in the LT are still low, so that target cells, $T(t)$, in the LT can be assumed to be at their pre-infection level. Viral loads in initial and systemic infections are not explicitly modeled but their parameters are absorbed into R_0^{muc} and R_0^{LT}, respectively. This standard approximation is based on the short lifetimes of virions (Perelson et al., 1996), which results in the virus load varying in time proportionally to the infected cell number. Thus, the model (Eqs. S1 or S2) can be simplified to exclude the variables $T(t)$ and $V(t)$ and keep p_{lat}, r, and R_0^{muc} and R_0^{LT} (Table S1) as the only input parameters. The simplest two-compartment version with discrete generations has the form:

$$I_{\text{muc}}(t+1) = Poisson\left[(1-p_{\text{lat}})R_0^{\text{muc}} I_{\text{muc}}(t)\right]$$
$$L_{\text{LT}}(t+1) = L_{\text{LT}}(t) + Poisson\left[p_{\text{lat}} R_0^{\text{muc}} I_{\text{muc}}(t) + p_{\text{lat}} R_0^{\text{LT}} I_{\text{LT}}(t)\right] - L_{\text{react}}$$
$$I_{\text{LT}}(t+1) = Poisson\left[(1-p_{\text{lat}})R_0^{\text{LT}} I_{\text{LT}}(t)\right] + L_{\text{react}}$$

I_{muc} and I_{LT} represent the numbers of actively infected (virus-producing) cells in initial and systemic infection, respectively, while L_{LT} represents the number of latently infected cells during systemic infection. Cells are latently infected with probability p_{lat}, and actively infected cells remain in their respective compartments and die. Time t =1, 2, 3, … is discrete and expressed in units of the infected-cell lifetime, $1/d_{\text{I}}$ = 1 day (Table S1), i.e., generations of actively infected cells are non-overlapping. This represents a generalized version of the Wright-Fisher process (unlike in the classical version, total subpopulations of cells are not fixed) (Nielsen and Slatkin, 2013).

Latently infected cells are resting memory CD4+ T cells [note that the opposite is not true: cells with resting markers can be actively infected as well presumably after relaxing into resting state (Zhang et al., 1999)], which can circulate freely between local mucosa and LT (Murphy, 2011). Hence, once formed in the mucosa, latent cells gain access to LT. The number of reactivating latent cells is $L_{\text{react}} = Poisson[r(t)\, L_{\text{LT}}(t)]$ where $r(t)$ is a time-dependent parameter describing the reactivation rate. We use an $r(t)$ that generates a 5-7 delay (Haase, 2011). Specifically, we assume that rare activation of latent cells occurs at a maximal rate r (Table S1) during initial mucosal infection (e.g., via T cell receptor when exposed to macrophages and dendritic cells, which express HIV peptides in MHC-II context and migrate from the mucosal infection site (Murphy, 2011)). After that, $r(t)$ was taken 0, for the short-term dynamics of several days [actually, it has finite but much smaller value < 10^{-3} which causes eventual cell

reactivation even after years of therapy (Hill et al., 2014; Rouzine et al., 2014), Table S1]. The general case with transfer of free virus or actively infected cells is considered separately below (*Section E).*

This simulation also explains the apparent conflict between the random nature of latent cell activation and the well-timed peak of viremia, which occurs at 10-12 days post transmission. According to our simulations (Fig S2F), although latent cells can be reactivated at any time with the same probability during the initial infection period, the cells that get activated later have the largest chance to survive until the influx of target cells occurs ($R_0 > 1$). Therefore, expansion of latent cells after transfer (red lines in Fig S2F) is almost exactly synced to the extinction of the initial active infection.

The coupled Wright-Fisher model was simulated in Matlab™ using the "broken-stick" method (Macarthur, 1957) to generate Poisson-distributed random numbers around their respective average values (Fig S1). Parameter values are given in Table S1.

Uncoupled compartments: deterministic approximation

To calculate the dependence of p_{estab} and I_0 on p in the deterministic approximation, we uncouple the mucosal and lymphoid sub-models and use Eqs S2 twice: once for the mucosal compartment in which $R_0^{\text{muc}} < 1$ (initial infection) and a second time for the lymphoid compartment in which $R_0^{\text{LT}} > 1$ (systemic infection). The mucosal model enables us to calculate the number of latently infected cells transferred, while the systemic infection model enables us to calculate the viral load and inoculation dose transferred to the next patient (see the main text for the references). This uncoupled approach does not explicitly consider the time delay (~ 6 days) between the start of mucosal and systemic infection (see a subsection below). We assume that the initial virus load in mucosa $V(0)$ which corresponds to an inoculum of $I_0 = I(1\text{day}) + L(1\text{day}) = 10\text{-}100$ initially infected cells (see *Section D*). For the lymphoid tissue (LT), we assume that a single reactivated cell seeds systemic infection: $I(0) = 1$. Other state variables are initially set to 0. We solved Eqs S2 numerically using ODE15s solver in the standard MATLAB package (Fig 2).

Net transmission rate as a function of latency probability (p): numeric and analytic (Fig 2D)

We assume that the average inoculum per unprotected sexual encounter I_0 is proportional to the average virus load $V(t)$

$$I_0 = const(p)\frac{1}{t_{\text{inf}}}\int_0^{t_{\text{inf}}} dt\, V(t) \tag{S4}$$

where $const(p)$ denotes a factor that does not depend on p, and $V(t)$ is calculated numerically from Eqs. S2. In this model, the peak and the steady state viremia are comparable, but the steady state is much longer (10 years on average) and hence dominates the integral in Eq. S4.

The net transmission probability $p_{\text{transmission}}$ can be approximated by

$$p_{\text{transmission}} \approx p_{\text{estab}} I_0 \tag{S5}$$

because $p_{\text{transmission}} << 1$ (Fraser et al., 2007). Here $p_{\text{estab}} = (L_{init}^{R0>1}/I_0)\, p_{\text{react}}$ is the probability of infection transfer to LT, including the fraction of latent cells formed in mucosa ($L_{init}^{R0>1}/I_0$) and the reactivation probability of a cell, p_{react}. Assuming that reactivation probability does not depend on p (we relax this assumption for the coupled model, see below), and $R_0^{\text{muc}} << 1$, p_{estab} is proportional to

the final level of latent cells formed in mucosa. Calculating $V(t)$ numerically from Eqs S2, normalized $p_{\text{transmission}}(p)$ from Eqs S4 and S5 is shown in Fig 2.

We can also obtain $p_{\text{transmission}}(p)$ analytically, as we do in the main text and repeat here. From Eqs S2, assuming $r << d_L$ (see *Section E*), steady-state viremia is given by $V = (d_T/k)\,[(1\text{-}p)\,R_0^{LT} - 1]$, where $R_0^{LT} = bkn/dc$. Hence, from Eq S1, we arrive at

$$I_0(p) \approx const \cdot \left[(1-p) R_0^{\text{LT}} - 1\right] \tag{S6}$$

Assuming $R_0^{muc} << 1$, the final level of latent cells formed in mucosa is approximately linearly proportional to p (Eq. S3). (Finite value R_0^{muc} creates a correction in p_{estab}; e.g., $R_0^{muc} = 0.25$ increases it by ~14%.) Hence, from Eq. S5, $p_{\text{transmission}}(p)$ is the quadratic dependence on p

$$p_{\text{transmission}}(p) \approx const \cdot p\left[(1-p) R_0^{\text{LT}} - 1\right] \tag{S7}$$

which is virtually identical to the numeric dependence (Fig 2). Thus, the latency probability has optimum at

$$p = p_{\text{opt}} = (1 - 1/\ R_0^{\text{LT}})/2 \tag{S8}$$

As we show below in *Section E*, same result for $p_{\text{transmission}}(p)$ follows approximately when transmission is dominated by the acute phase of infection and the model includes the immune response against HIV and in some other cases as well. In general, the dependence of transmission rate on p is surprisingly robust to model variations (*Section E*).

C. Extended model of systemic infection in lymphoid tissue ($R_0^{LT} > 1$) including the adaptive immune response

The basic model (Fig 1B, Fig 2) is the simplest model in the literature that explores the idea of latency-dependent transmission. It predicts the high early incidence of latency observed in cell culture and 3 days postinfection in experimental animals. However, the simple model also has limitations: for the ensuing chronic infection, it predicts an unrealistically high latent compartment. To interpret the low reservoir size observed in chronic infection, we introduce an extended model including immune response. Importantly, the central result of a high $p_{\text{opt}}(0)$ is robust to this and other model changes (Section E).

The critical importance of the cytotoxic immune response for the HIV-1 dynamics is evident from experiments on CD8 T-cell depletion in monkeys (Jin et al., 1999; Schmitz et al., 1999) and rapid genetic variation in CD8 T-cell epitopes. To test the robustness of the results to a model, and to introduce the time variation of latency parameters *in vivo*, we expanded the standard model to include an antigen-specific CD8 T-cell response (Davenport et al., 2004; De Boer, 2007; De Boer and Perelson, 1998), as follows:

Uninfected target cells $$\frac{dT}{dt} = \underbrace{b}_{\text{replenishment}} - \underbrace{d_T T}_{\text{naturaldeath}} - \underbrace{kVT}_{\text{infection}}$$

Infected cells in eclipse phase $$\frac{dI_E}{dt} = \underbrace{(1-p)kVT}_{\text{active infection}} - \underbrace{d_{IE} I_E}_{\text{eclipse phase ends}} + \underbrace{rL}_{\text{reactivation}}$$

Actively infected cells $$\frac{dI}{dt} = \underbrace{d_{IE} I_E}_{\text{from eclipse phase}} - \underbrace{d_I (1 + E/E_0) I}_{\text{death}}$$

Latently infected cells $$\frac{dL}{dt} = \underbrace{pkVT}_{\text{latent infection}} - \underbrace{d_L L}_{\text{death}} - \underbrace{rL}_{\text{reactivation}} \qquad \text{(S9)}$$

Free virus $$\frac{dV}{dt} = \underbrace{n d_P I}_{\text{production}} - \underbrace{cV}_{\text{clearance}}$$

Effector immune cells $$\frac{dE}{dt} = \underbrace{aE_N I/(I + I_{av})}_{\text{from naive cells}} + \underbrace{aEI/(I + I_{av})}_{\text{antigen-induced proliferation}} - \underbrace{d_E E}_{\text{death}}$$

Naive precursors of immune cells $$\frac{dE_N}{dt} = -\underbrace{aE_N I/(I + I_{av})}_{\text{activation by antigen}}$$

All new parameters listed in the extended immune model above are described in Table S2 (Table S1 describes the parameters inherited from the Basic Model). Compared to the basic model (Eqs. S2), which assumes that the death rate of virus-producing cells d_I (caused by viral products) is constant, in the extended model, effector CD8 T cells (E) can also kill these cells (the term E/E_0 in line 3 in Eqs. S9). Also added are two new equations, describing the activation of naive CD8 T cells (E_N) by infected cells and the expansion and the death of effector cells. [As is usually done with models including immune clearance, we also included an eclipse phase, I_E, to limit the decay rate of actively infected cells at finite value less than 1/day and thus prevent an artifact of their precipitous drop and giant oscillations (De Boer, 2007; Rouzine et al., 2006; Sergeev et al., 2010b).] Eqs S9 represent one of the simplest models of the CD8 T cells response. More complex models involve helper T cell dependent activation (Sergeev et al., 2010a; Sergeev et al., 2010b) or multiple CTL clones recognizing different epitopes (Althaus and De Boer, 2008). Yet, even this simple model is sufficient for our aim, because the effects of the CD8 response are sufficient to capture the robustness of the results (next subsection). Further attempts at model simplification produce unobserved effects on dynamics (Fig S3B, C).

In Eqs. S9, we postulate time variation of latency parameters driven by cytokine activation of target cells $T(t)$ and latently infected cells $L(t)$ by immune cells $E(t)$, as follows (Fig. 3)

$$p(E) = p(0)E_{0L}/(E_{0L}+E), \qquad r(E) = r(0) + d_I E/(E_{0L}+E) \qquad \text{(S10)}$$

Here $E = E_{0L}$ is the characteristic number of immune cells at which half of infected cells receive a cytokine signal above an activation threshold, and the initial values $p(0) \sim 0.5$ (Eq S8) and $r(0) = 10^{-4}$-10^{-3} (Table S1) correspond to the absence of the HIV-specific immune response.

The dynamics of cell compartments calculated from Eqs. S9 and S10 is shown in Fig. 4A. Model parameters are described in Table S2 and Section D below. In the beginning of systemic expansion, $p = p(0) \sim 0.5$ and $r = r(0) << 1$ (Fig 4B), so that the latent cell count rises to high levels. The immune

response after the viremia peak activates target cells, which increases r and decreases p by several orders of magnitude and thus depletes the latent reservoir. The latently infected cell subpopulation (L) saturates at a low level, determined by the dynamic balance between new infections and activation of latently infected cells (Fig 4A). Latently infected cells (i.e., cells in the "off" state of virus transcription) are being constantly produced by new infections and immediately activated back to the "on" state (I). Thus, the size of latent compartment remains dynamically coupled to the actively infected compartment.

The onset of ART suppresses new infections and results in a decline of all infected cells, both virus-producing and latent because they are coupled dynamically (Fig 4A). The decline of the latent reservoir occurs at the rate of cell activation, $r(t)$. At the onset of ART, this rate is very high, and latent compartment decays very rapidly. Because the depletion of antigen causes immune cell population to contract (E), the reactivation rate of latently infected cells $r(t)$ soon returns back to its low background level $r(0)$ (Table S1). Therefore, any further decay of the reservoir occurs extremely slowly (Fig 4A).

D. Model parameters and parameter sensitivity analysis

HIV demographics in early mucosa ($R_0^{muc} < 1$)

The assumption $R_0^{muc} < 1$ for the initial entry site of HIV is critical. That assumption--and our entire model-- is built on the facts described in Haase, 2011 (Miller et al 2005, Li et al 2009). These results and the supporting data by *in situ* hybridization are consistent with our assumption that reproductive number in mucosa (R_0) is below 1 through day 5 of infection. Following day 6 and on, a local RNA expansion has been observed, roughly synchronous with the RNA expansion in genital and distal lymph nodes, which demonstrates that $R_0 > 1$ from that time on (Miller et al 2005, Fig 1A-C). Thus, our target-poor $R_0^{LT} < 1$ compartment (termed "mucosa" for brevity) includes early mucosa, while our target-rich $R_0^{LT} > 1$ compartment (termed "lymphoid tissue") (Fig 1B) includes late mucosa, genital and distal lymph nodes, GALT, spleen, and other major organs of HIV replication.

Miller et al (2005) inoculated female macaques with a 1ml viral dose with the high concentration of 10^9 RNA copy/ml (during a typical transmission, it is $\sim 10^5$ RNA copy/ml sperm). The authors failed to detect any consistent evidence for active infection in the interval 0-5 days post-infection (see their Fig 1A). No SIV RNA was consistently detected in vagina or endocervix until day 6 (except for residual inoculum on day 1). Using a more sensitive assay of in situ hybridization capable of detecting single SIV RNA+ cells, a founder population of 40-50 actively infected (SIV RNA+) cells was discovered in a single animal, out of 14 total animals necropsied before day 6 (Miller et al 2005, Fig 3 and the text). A more careful and broad search (Li et al 2009) discovered such founder populations (on day 4) in 9 animals; unfortunately, these authors did not specify the total number of animals scanned, so that we cannot use these data. We emphasize that natural viral transmission occurs at 10,000-fold lower viral concentrations, making founder populations even less frequent. Finally, we conclude that active viral replication is probably extinct in this time window, and, therefore, the reproduction ratio during this time R_0 is below 1.

Estimate of I_0 from the count of HIV DNA+ cells in early mucosa

Thus, data show no active viral replication (no SIV RNA+ cells) in most animals until day 6. At the same time, recent data demonstrate the existence of a large latent compartment. The observed frequency of SIV DNA+ cells inside a mesenteric lymph node on day 3 postinfection is 200 per 10^6 CD4+ cells (Whitney et al, 2014, see their Fig S5). From this value, assuming 300 CD4 cell/μg (Zhang et al., 1998) in a 2g node [a node weighs 10 mg in a mouse (Kim et al., 2008) whose body weight is 200-fold smaller than that of a rhesus macaque], we arrive at the latent reservoir of $L_{init}^{R_0>1} = 10^5$ SIV DNA+ cells (assuming that all DNA+ cells are latently infected and do not harbor defective proviruses). These were inoculated by high dose of virus, 1ml supernatant with the concentration 10^9 SIV RNA/ml. For the 10^4-fold less concentrated virus during typical transmission, we estimate $L_{init}^{R_0>1} \sim 10$ DNA+ cells, which corresponds to inoculum $I_0 = 20$-30 if $p = 0.5$ chosen in our simulations (Figs 2 and S2).

Parameter choice and sensitivity in target-rich compartment ($R_0^{LT} > 1$)

For the basic model in Figs 1 and 2, we used standard parameters from the literature, as cited in Table S1. Importantly, our predicted result for p_{opt} is affected by only one composite parameter, R_0^{LT}. Parameter sensitivity analysis (Fig. 2) to verify robustness of the optimum in net-transmission rate was carried out by varying all model parameters via R_0^{LT}. R_0^{LT} was varied within the measured range in patients or in rhesus macaques ($R_0^{LT} = 5$ to $R_0^{LT} = 20$) (Nowak et al., 1997). Other parameters estimated from the literature (such as death rates d_T, d_I, see Table S1) affect the dynamics of the acute viremia peak, but not p_{opt}.

Definition of parameters is given in Table S2. Out of the seven parameters four (d_E, E_0, E_{0L}, I_0) are adjusted to fit the four experimental plateaus in patients (Fig 4), and three are fixed and cited from the literature. Specifically, E_0, I_0, and E_{0L} are estimated from the predicted steady-state levels of E, I, and L, compared to their experimental estimates (Fig. 4): $E = 10^9$ cells (Ogg et al., 1998) (Turnbull et al., 2009), $I = 10^8$ cells (Haase, 1999), and $L = 10^6$ cells (Chun et al., 1997). d_E is estimated to fit L under ART: $L = 10^5$ cells (Finzi et al., 1997). The total cell counts are assumed $T(0) = b/d_T = 2 \cdot 10^{11}$ for both $CD8^+$ T and $CD4^+$ T cells (Murphy, 2011).

E. Robustness to the variations of the individual-host model and epidemiological factors

To test the robustness of the $p_{opt} \sim 0.5$ prediction to changes in model structure, we modified the model architecture in four ways: (i) extended the model to include an immune response, (ii) altered the relative contribution of acute versus chronic stages to transmission potential, (iii) examined a non-linear dependence of the transmission rate on viremia, and (iv) altered the role of actively producing cells in the transfer between mucosa and the lymph. As we show below, although viral dynamics is generally sensitive to many factors and parameters, and viral transmission may occur in different phases of HIV, the normalized dependence of the transmission probability on p is surprisingly robust (or skewed towards even larger p). As long as model parameters stay within their order of magnitude in Table S1, all the variation occurs within the constant prefactor in Eq. S7 and does not alter the dependence on $p(0)$.

Limitations of the basic model (Eqs S2): high latent cells and sensitivity to r/d_L

In the main text derivation, we assumed $r << d_L$. Actually, the existing estimates of r and d_L are within the same range, so that both $r < d_L$ and $r > d_L$ are possible (Table S1). In the opposite extreme scenario, $r >> d_L$, the virus load increases by a factor of 1-p (Eqs S2):

$$V = \frac{d_T}{k}\left[R_0^{LT}\frac{1-p+\frac{r}{d_L}}{1+\frac{r}{d_L}}-1\right]$$

Thus, the virus load in steady state V is sensitive to r/d_L even when both r and d_L are very small, $< 10^{-3}$. Importantly, the optimal value p_{opt} stays large regardless. In the case $r >> d_L$, Eq S6 for $I_0(p)$ loses factor 1-p before R_0^{LT}, so that $I_0(p)$ calculated in the basic model becomes independent on p. As a result of the change, the optimum value for the initial p_{opt} becomes even higher than 0.5 (Eq S8). Thus, p_{opt} is large regardless of the ratio r/d_L.

The sensitivity of V to r/d_L is an artifact of the basic model (Eqs S2), which, as it is well known, cannot be used to predict the peak to steady state ratio. Another artifact is a very high level predicted for latent cells L. The steady state values for T and L are:

$$T = \frac{b_{LT}}{R_0^{LT}d_T}\frac{1+\frac{r}{d_L}}{1-p+\frac{r}{d_L}}$$

$$L = \frac{b_{LT}p}{(r+d_L)R_0^{LT}}p\left[R_0^{LT}-\frac{1+\frac{r}{d_L}}{1-p+\frac{r}{d_L}}\right]$$

Therefore, at $p \sim 0.5$, the ratio L/T is as large as $R_0^{LT}d_T/(d_L+r) \sim 10^3$-$10^4$ (see parameters in Table S1).

The unrealistically high L (and V) predicted by the basic model (Eqs S2) indicate that the basic model, although capable of interpreting high levels of latency in early infection, is not sufficient for interpreting low levels of latency (and other parameters) in chronic infection (Fig 4A), and must be extended to a more realistic model including an immune response (Eqs. S9-S10). Importantly, as we next show, the high initial latency probability $p_{opt}(0) \sim 0.5$ is robust to this (and other) model modifications.

Factors affecting the transmitted dose

So far, we have used Eq. S4, for I_0 assuming linear dependence on average virus load. The actual number of infected units transmitted to an average individual is a more complicated quantity that depends on epidemiological factors, such as a risk group (frequency of sex contacts) and variation of host's infectivity with the stage of infection (Baggaley et al., 2006). For our aim of estimating p_{opt}, we consider only high-risk groups of hosts, which are expected to dominate the HIV spread and evolution. Further, HIV infection can be split into a highly infectious acute stage including viremia peak (~ 1-2 months) and a less infectious but much longer chronic stage (~100 months). In high-risk groups of humans, half of transmissions occur early during the acute stage (Fraser et al., 2007; Hollingsworth et al., 2008; Lewis et al., 2008; Wawer et al., 2005); we expect a similar contribution from the acute phase

in a natural host population. Note that, since a natural host does not develop AIDS, the highly infectious AIDS stage is absent.

Basic model and acute-stage transmission

Consider now the contribution of the acute phase to transmission (the time integral in Eq. S4) as a function of p. The initial expansion slope $d\ln I(t)/dt$ is equal to the initial reproduction number $(1\text{-}p)R_0^{\mathrm{LT}}$ (which is smaller than the raw value in the absence of latency (R_0^{LT}) due to diversion of infected cells from active viral production). As the population of infected cells expands, uninfected target cells (T) are depleted proportional to the viral load (Fig S3A, blue dashed line). The expansion is checked and the viral load reaches its maximum when the reproduction ratio is decreased from the initial value $(1\text{-}p)R_0^{\mathrm{LT}}$ to 1 due to depletion of T. Thus, to stop virus expansion and reach the peak, the target cells must be depleted by a factor of $(1\text{-}p)R_0^{\mathrm{LT}}$. Since depletion is proportional to the virus load, the virus peak height must be proportional to $1–p$. Using this proportionality in Eq. S4 contributed mostly from the peak, we again arrive at Eq. S6 for I_0, and at the same result for p_{opt} as for the chronic-phase transmission (Eq S8).

Immune response and acute-stage transmission

The presence of CD8 T cells in the extended model (given by Eqs. S9-S10), which become prominent after the viremia peak, decreases the steady-state virus load further than in the base model given by Eqs. S2 (Fig S3A). Yet, the region of the viremia peak (which determines I_0 in acute-stage transfer from Eq S4) do not change much. As in the absence of the immune response, the height of the infection peak is limited by the depletion of target cells, which occurs before the immune cells rise to prominence. The shape of the viremia peak near its maximum is also fairly robust. Indeed, the decay slope after the maximum is determined by lifetime of the eclipse phase cells (I_E in Eqs. S9) (Klenerman et al., 1996a; Rouzine et al., 2006). Assuming, again, the dominant role of the acute stage in transmission, we predict almost the same result for the transmission rate as a function of p as for the model without an immune response. This is confirmed by numerical simulation. To obtain the result shown in Fig 4B, we compute Eqs. S9 and S10 numerically and, assuming that most transmission occurs during the viremia peak, evaluate the time integral in Eq. S4 over the interval of 20 days centered at the peak. Indeed, numerical results for $p_{\mathrm{transmission}}(p)$ in Fig 4B (with immune response) are similar to that in Fig 2D (no immune response).

Immune response and mixed acute-chronic-stage transmission

We assume now that 50% transmission events take place during the peak and 50% during the chronic phase (Fraser et al., 2007). We also assume that p [pre-immune-response value $p \equiv p(0)$] is at the evolved optimum, p_{opt}. The immune response is present (Eqs. S9-S10). While the peak part of the integral in Eq. S4 is still proportional to $1\text{-}p$, the steady-state part does not depend on p at all, because virus is pinned near the CTL avidity threshold, $I \sim I_0$. Hence, the average inoculum is given by the sum of the peak part and the steady state part:

$$\left[R_0^{\mathrm{LT}}(1-p)-1\right]/2+\left[R_0^{\mathrm{LT}}(1-p_{\mathrm{opt}})-1\right]/2 \qquad \text{(S6)}$$

The first term due to peak is from Eq. S6 and the second term due to steady state (which is constant in p) is calculated from the condition that both terms are equal at $p = p_{\mathrm{opt}}$. As one derives easily, the best net transmission rate given by the product $p\cdot I_{\mathrm{inoculum}}(p)$ will be attained at $p = p_{\mathrm{opt}} = (2/3)(1\text{-}1/R_0^{\mathrm{LT}})$, larger

than the main text result $p_{opt} = (1/2)(1-1/R_0^{LT})$. Thus, inclusion of the two infection stages into transmission in the presence of the immune response will only increase the optimum above ~ 0.5. Once again, the prediction of a large optimal latency probability in the beginning of infection $p = p(0) > 0.5$ remains.

Non-linear dependence of the transmission rate on the viremia

Another possible factor is that the probability of transmission is not linear in the viral and infected cell counts, as is assumed in Eq. S4, but saturates at high viremia levels (as in HIV-status discordant couples) (Fraser et al., 2007). Transmission rates are higher in high-risk groups of individuals (May, 2004). In any case, saturation of the transmission rate would also only increase p_{opt}, because a slower-than-linear increase of the transmission rate with viremia would translate to a slower decrease of the transmission rate with p, as compared to the basic model result (Eq S7). This would reduce the cost of latency. Hence, again, p_{opt} will only increase. Thus, as long as we abide by the central hypothesis of the present work that latently infected cells seed systemic HIV infections, the prediction of a large optimal p remains. Now we have to verify what happens if we relax the central hypothesis.

Transmission in the presence of non-latent virus transfer (Fig 2E, Fig S2J)

Before we focused on the case when only latent cells can seed systemic infections. Now we relax the central hypothesis and assume that the viral progeny of actively infecting cells (including CD4 T cells and dendritic cells) can also seed systemic infections. In this case, the generalized expression for the effective transmission rate takes a form (both for the basic model and for the extended immune response model):

$$p_{transmission} = p_{estab}\, I_0 \qquad \text{(S5')}$$

$$I_0(p) = const\,(p)\left(1 - p - \frac{1}{R_{0LT}}\right), \qquad \text{(S6')}$$

$$p_{estab}(p) = Const(p)\,[(1-p)f_{act} + (1-f_{act})p\,] \qquad \text{(S11)}$$

Here p is the pre-immune response latency probability, $p \equiv p(0)$. The new parameter f_{act} is defined to be the fraction of systemic infections due to non-latent routes when $p = 0.5$. The new establishment probability $p_{estab}(p)$ in Eq. S5 includes the probability that an actively infected cell seeds systemic infection (proportional to the probability of active infection $1-p$). Notably, the right-hand side of Eq. S11 is just one parameterization—it can more generally be re-parameterized in the linear form: $A+Bp$.

At a fixed f_{act}, we then calculate the new optimum in p

$$p_{opt} = \frac{1}{2}\left(1 - \frac{1}{R_{0LT}} - \frac{f_{act}}{1-2f_{act}}\right) \qquad \text{(S12)}$$

and note it to be less than the result for p_{opt} obtained at $f_{act} = 0$. Eq. S12 is valid for $f_{act} \le (1-1/R_{0LT})/[1+2(1-1/R_{0LT})]$ to ensure $p_{opt} \ge 0$.

Critically, the predicted and experimentally relevant fraction of active-cell transfer denoted f_{nonlat} is at $p = p_{opt}$.

$$f_{nonlat} = \frac{(1-p_{opt})f_{act}}{(1-p_{opt})f_{act}+p_{opt}(1-f_{act})} \quad \text{(S13)}$$

which is obtained as the relative weight of the first term in Eq. S11. It is larger than the raw value f_{act}. As f_{act} increases to $\approx \frac{1}{3}$, the value of p_{opt} vanishes, and transfer switches 100% to the active-cell component, f_{nonlat}=1. Until then, the transfer is mixed, $f_{nonlat} < 1$, and latency probability has non-zero optimum. For example, for 90% and 10% split of the transmission role between active and latent cells, we have $p_{\text{opt}} \sim 0.05$. Thus, for latency to become useless to the virus (i.e., $p_{\text{opt}} = 0$), actively infected cells (or infected dendritic cells) must completely dominate the seeding of systemic infection. Otherwise, the possibility of latency is beneficial to the virus.

Dependence of establishment probability* (p_{estab}) and *reactivation probability* (p_{react}) *on* p_{lat} *(Fig S2)

In the derivation of $p_{\text{transmission}}(p)$ of the main text, we assumed that the probability of reactivation of a latent cell in LT (p_{react}) does not depend on the probability of latency, p (Fig. 2C). This is a natural assumption for the basic decoupled ODE model on which this derivation is based (Eqs S2, Fig. 2), which does not include the rate of reactivation of latent cells explicitly (the initial condition in the LT is one reactivated cell).

To explicitly test the case in which p_{react} depends on p, we also conducted a Wright-Fisher simulation of the stochastic coupled model described above in *Section B*. In this coupled simulation, the value of p_{react} decreases with p by 60% from its maximum value at $p = 0$ (Fig S2G). The reason for the decrease is our assumption that activation of latent cells occurs during active mucosal infection; as p increases, active infection starts at lower levels, and becomes extinct faster, shrinking the duration of activation. Critically, even with this extreme assumption, p_{opt} only changes from 0.45 (Fig 2D) to 0.35 (Fig S2H). Thus, the key result that p_{opt} is large remains robust.

The peak of latent cells is sensitive to the details of latency control by the immune response

The size of the peak of latent cells in acute infection in the LT [the purple curve $L(t)$ in Fig. 5A] is very sensitive to the details and parameters through which CD8+ T cells control latently infected cells (these parameters are currently determined by Eqs S10). Specifically, the sensitive parameters include: (i) the minimal possible value of $p(E)$ (currently, set to 0), (ii) the maximum value of $r(E)$ (currently, set to d_{I} = 1/day), (iii) the characteristic level of $E = E_{0\text{L}}$, at which CD8 T cells reactivate latent cells and decrease $p(E)$ (currently $E_{0\text{L}} = 4{\cdot}10^6$ cells, but it may be much lower), and (iv) the heterogeneity of the latent cell population due to variation in the HIV gene integration site (Dar et al., 2012), which includes a fraction of latent cells resisting activation (Ho et al., 2013) (currently not included). Data for these details are currently incomplete or absent. Therefore, at this time, we cannot make a quantitative prediction regarding the size (or even the existence) of the $L(t)$ peak during acute infection.

Simplified immune models fail to predict the realistic viral dynamics (Fig S3B-D)

To test whether the extended immune model (Eqs S9, S10) could be simplified further, we investigated two representative model reductions:

No immune eclipse cells. In the first simplified model (Fig 3B,C), we eliminated cells in the eclipse phase, I_{E}, replacing the first 3 equations in Eqs S9 with 2 equations:

Uninfected target cells $$\frac{dT}{dt} = \underbrace{b}_{\text{replenishment}} - \underbrace{d_T T}_{\text{naturaldeath}} - \underbrace{kVT}_{\text{infection}}$$

Actively infected cells $$\frac{dI}{dt} = \underbrace{kVT}_{\text{infection}} - \underbrace{d_I(1+E/E_0)I}_{\text{death}}$$

Numerical simulations of this simplified model show a precipitous ~1 log drop of infected cells and viremia in less than 1 day at the onset of ART (Figs. S3B, C). In contrast, in patients, viremia decays at the rate of ~1/day (see Fig. 4A, blowup).

No target cell depletion. In the second simplified version (Fig. S3D), we neglected the depletion of target cells. The first equation in Eqs S9 was eliminated, and T was fixed at its predicted value in uninfected patients: $T = b/d_T$. Numerical simulations show a giant peak of infected cells during acute infection overshooting the uninfected T level and strong oscillations (Fig. S3D). These features are not observed in patients or in experimental animals who show dynamics similar to those in Figure 4A.

SUPPLEMENTAL REFERENCES

Althaus, C.L., and De Boer, R.J. (2008). Dynamics of immune escape during HIV/SIV infection. PLoS Comp Bio *4*, e1000103.

Baggaley, R.F., Garnett, G.P., and Ferguson, N.M. (2006). Modelling the impact of antiretroviral use in resource-poor settings. PLoS Med *3*, e124.

Brandin, E., Thorstensson, R., Bonhoeffer, S., and Albert, J. (2006). Rapid viral decay in simian immunodeficiency virus-infected macaques receiving quadruple antiretroviral therapy. J Virol *80*, 9861-9864.

Chun, T.W., Carruth, L., Finzi, D., Shen, X., DiGiuseppe, J.A., Taylor, H., Hermankova, M., Chadwick, K., Margolick, J., Quinn, T.C.*, et al.* (1997). Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. Nature *387*, 183-188.

Dar, R.D., Razooky, B.S., Singh, A., Trimeloni, T.V., McCollum, J.M., Cox, C.D., Simpson, M.L., and Weinberger, L.S. (2012). Transcriptional burst frequency and burst size are equally modulated across the human genome. Proc Nat Acad of Sci U S A *109*, 17454-17459.

Davenport, M.P., Ribeiro, R.M., and Perelson, A.S. (2004). Kinetics of virus-specific CD8+ T cells and the control of human immunodeficiency virus infection. J Virol *78*, 10096-10103.

De Boer, R.J. (2007). Understanding the failure of CD8+ T-cell vaccination against simian/human immunodeficiency virus. J Virol *81*, 2838-2848.

De Boer, R.J., Homann, D., and Perelson, A.S. (2003). Different dynamics of CD4 and CD8 T cell responses during and after acute lymphocytic choriomeningitis virus infection. J Immunol *171*, 3928-3935.

De Boer, R.J., and Perelson, A.S. (1998). Target cell limited and immune control models of HIV infection: a comparison. J Theor Biol *190*, 201-214.

Finzi, D., Hermankova, M., Pierson, T., Carruth, L.M., Buck, C., Chaisson, R.E., Quinn, T.C., Chadwick, K., Margolick, J., Brookmeyer, R.*, et al.* (1997). Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. Science *278*, 1295-1300.

Fraser, C., Hollingsworth, T.D., Chapman, R., de Wolf, F., and Hanage, W.P. (2007). Variation in HIV-1 set-point viral load: epidemiological analysis and an evolutionary hypothesis. Proc Nat Acad Sci U S A *104*, 17441-17446.

Grimmett, G., and Stirzaker, D. (2001). Probability and random processes, 3rd edn (Oxford ; New York, Oxford University Press).

Haase, A.T. (1999). Population biology of HIV-1 infection: viral and CD4+ T cell demographics and dynamics in lymphatic tissues. Annu Rev Immunol *17*, 625-656.

Haase, A.T. (2011). Early events in sexual transmission of HIV and SIV and opportunities for interventions. Annu Rev Med *62*, 127-139.

Hill, A.L., Rosenbloom, D.I., Fu, F., Nowak, M.A., and Siliciano, R.F. (2014). Predicting the outcomes of treatment to eradicate the latent reservoir for HIV-1. Proc Nat Acad Sci U S A *111*, 13475-13480.

Ho, Y.C., Shan, L., Hosmane, N.N., Wang, J., Laskey, S.B., Rosenbloom, D.I., Lai, J., Blankson, J.N., Siliciano, J.D., and Siliciano, R.F. (2013). Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. Cell *155*, 540-551.

Hollingsworth, T.D., Anderson, R.M., and Fraser, C. (2008). HIV-1 transmission, by stage of infection. J Infec Dis *198*, 687-693.

Jin, X., Bauer, D.E., Tuttleton, S.E., Lewin, S., Gettie, A., Blanchard, J., Irwin, C.E., Safrit, J.T., Mittler, J., Weinberger, L.*, et al.* (1999). Dramatic rise in plasma viremia after CD8(+) T cell depletion in simian immunodeficiency virus-infected macaques. J Exp Med *189*, 991-998.

Kim, C.S., Lee, S.C., Kim, Y.M., Kim, B.S., Choi, H.S., Kawada, T., Kwon, B.S., and Yu, R. (2008). Visceral fat accumulation induced by a high-fat diet causes the atrophy of mesenteric lymph nodes in obese mice. Obesity (Silver Spring) *16*, 1261-1269.

Klatt, N.R., Shudo, E., Ortiz, A.M., Engram, J.C., Paiardini, M., Lawson, B., Miller, M.D., Else, J., Pandrea, I., Estes, J.D.*, et al.* (2010). CD8+ lymphocytes control viral replication in SIVmac239-infected rhesus macaques without decreasing the lifespan of productively infected cells. PLoS Pathog *6*, e1000747.

Klenerman, P., Phillips, R.E., Rinaldo, C.R., Wahl, L.M., Ogg, G., May, R.M., McMichael, A.J., and Nowak, M. (1996a). Cytotoxic T lymphocytes and viral turnover in HIV type 1 infection. Proc Natl Acad Sci USA *93*, 15323-15328.

Klenerman, P., Phillips, R.E., Rinaldo, C.R., Wahl, L.M., Ogg, G., May, R.M., McMichael, A.J., and Nowak, M.A. (1996b). Cytotoxic T lymphocytes and viral turnover in HIV type 1 infection. Proc Natl Acad Sci U S A *93*, 15323-15328.

Lewis, F., Hughes, G.J., Rambaut, A., Pozniak, A., and Leigh Brown, A.J. (2008). Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS Med *5*, e50.

Li, Q., Duan, L., Estes, J.D., Ma, Z.M., Rourke, T., Wang, Y., Reilly, C., Carlis, J., Miller, C.J., and Haase, A.T. (2005). Peak SIV replication in resting memory CD4+ T cells depletes gut lamina propria CD4+ T cells. Nature *434*, 1148-1152.

Macarthur, R.H. (1957). On the Relative Abundance of Bird Species. Proc Natl Acad Sci U S A *43*, 293-295.

Markowitz, M., Louie, M., Hurley, A., Sun, E., Di Mascio, M., Perelson, A.S., and Ho, D.D. (2003). 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 *77*, 5037-5038.

May, R.M. (2004). Uses and abuses of mathematics in biology. Science *303*, 790-793.

Miller, C.J., Li, Q., Abel, K., Kim, E.Y., Ma, Z.M., Wietgrefe, S., La Franco-Scheuch, L., Compton, L., Duan, L., Shore, M.D.*, et al.* (2005). Propagation and dissemination of infection after vaginal transmission of simian immunodeficiency virus. J Virol *79*, 9217-9227.

Murphy, K. (2011). Janeway's Immunobiology, Eighth Edition (London and New York, Garland Science).

Nielsen, R., and Slatkin, M. (2013). An introduction to population genetics: Theory and applications, 1st edn (Sinaur Associates, Inc).

Nowak, M.A., Lloyd, A.L., Vasquez, G.M., Wiltrout, T.A., Wahl, L.M., Bischofberger, N., Williams, J., Kinter, A., Fauci, A.S., Hirsch, V.M.*, et al.* (1997). Viral dynamics of primary viremia and antiretroviral therapy in simian immunodeficiency virus infection. J Virol *71*, 7518-7525.

Ogg, G.S., Jin, X., Bonhoeffer, S., Dunbar, P.R., Nowak, M.A., Monard, S., Segal, J.P., Cao, Y., Rowland-Jones, S.L., Cerundolo, V.*, et al.* (1998). Quantitation of HIV-1-specific cytotoxic T lymphocytes and plasma load of viral RNA. Science *279*, 2103-2106.

Pearson, J.E., Krapivsky, P., and Perelson, A.S. (2011). Stochastic theory of early viral infection: continuous versus burst production of virions. PLoS Comput Biol *7*, e1001058.

Perelson, A.S., Neumann, A.U., Markowitz, M., Leonard, J.M., and Ho, D.D. (1996). HIV-1 dynamics in vivo: virion clearance rate, infected cell life span, and viral generation time. Science *271*, 1582-1586.

Rong, L., and Perelson, A.S. (2009a). Modeling HIV persistence, the latent reservoir, and viral blips. J Theor Biol *260*, 308-331.

Rong, L., and Perelson, A.S. (2009b). Modeling latently infected cell activation: viral and latent reservoir persistence, and viral blips in HIV-infected patients on potent therapy. PLoS Comp Bio *5*, e1000533.

Rouzine, I.M., Razooky, B.S., and Weinberger, L.S. (2014). Stochastic variability in HIV affects viral eradication. Proc Natl Acad Sci U S A *111*, 13251-13252.

Rouzine, I.M., Sergeev, R.A., and Glushtsov, A.I. (2006). Two types of cytotoxic lymphocyte regulation explain kinetics of immune response to human immunodefiency virus. Proc Natl Acad Sci U S A *103*, 666-671.

Schmitz, J.E., Kuroda, M.J., Santra, S., Sasseville, V.G., Simon, M.A., Lifton, M.A., Racz, P., Tenner-Racz, K., Dalesandro, M., Scallon, B.J.*, et al.* (1999). Control of viremia in simian immunodeficiency virus infection by CD8+ lymphocytes. Science *283*, 857-860.

Sedaghat, A.R., Siliciano, J.D., Brennan, T.P., Wilke, C.O., and Siliciano, R.F. (2007). Limits on replenishment of the resting CD4+ T cell reservoir for HIV in patients on HAART. PLoS Pathog *3*, e122.

Sedaghat, A.R., Siliciano, R.F., and Wilke, C.O. (2008). Low-level HIV-1 replication and the dynamics of the resting CD4+ T cell reservoir for HIV-1 in the setting of HAART. BMC Infect Dis *8*, 2.

Sergeev, R.A., Batorsky, R.E., Coffin, J.M., and Rouzine, I.M. (2010a). Interpreting the effect of vaccination on steady state infection in animals challenged with Simian immunodeficiency virus. J Theor Biol *263*, 385-392.

Sergeev, R.A., Batorsky, R.E., and Rouzine, I.M. (2010b). Model with two types of CTL regulation and experiments on CTL dynamics. J Theor Biol *263*, 369-384.

Stafford, M.A., Corey, L., Cao, Y., Daar, E.S., Ho, D.D., and Perelson, A.S. (2000). Modeling plasma virus concentration during primary HIV infection. J Theoretical Biol *203*, 285-301.

Turnbull, E.L., Wong, M., Wang, S., Wei, X., Jones, N.A., Conrod, K.E., Aldam, D., Turner, J., Pellegrino, P., Keele, B.F.*, et al.* (2009). Kinetics of expansion of epitope-specific T cell responses during primary HIV-1 infection. J Immunol *182*, 7131-7145.

Wawer, M.J., Gray, R.H., Sewankambo, N.K., Serwadda, D., Li, X., Laeyendecker, O., Kiwanuka, N., Kigozi, G., Kiddugavu, M., Lutalo, T.*, et al.* (2005). Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in Rakai, Uganda. J Infect Dis *191*, 1403-1409.

Zhang, Z., Schuler, T., Zupancic, M., Wietgrefe, S., Staskus, K.A., Reimann, K.A., Reinhart, T.A., Rogan, M., Cavert, W., Miller, C.J.*, et al.* (1999). Sexual transmission and propagation of SIV and HIV in resting and activated CD4+ T cells. Science *286*, 1353-1357.

Zhang, Z.Q., Notermans, D.W., Sedgewick, G., Cavert, W., Wietgrefe, S., Zupancic, M., Gebhard, K., Henry, K., Boies, L., Chen, Z.*, et al.* (1998). Kinetics of CD4+ T cell repopulation of lymphoid tissues after treatment of HIV-1 infection. Proc Natl Acad Sci U S A *95*, 1154-1159.